Week-01 : NumPy

1. Numpy

This is a quick overview of arrays in NumPy. It demonstrates how n-dimensional (n>=2) arrays are represented and can be manipulated. In particular, if you don't know how to apply common functions to n-dimensional arrays (without using for-loops), or if you want to understand axis and shape properties for n-dimensional arrays.

Learning Objectives

You should be able to:

- Understand the difference between one-, two- and n-dimensional arrays in NumPy;
- Understand how to apply some linear algebra operations to n-dimensional arrays without using for-loops;
- Understand axis and shape properties for n-dimensional arrays.

The Basics

NumPy's main object is the homogeneous multidimensional array. It is a table of elements (usually numbers), all of the same type, indexed by a tuple of non-negative integers. In NumPy dimensions are called axes.

For example, the array for the coordinates of a point in 3D space, [1, 2, 1], has one axis. That axis has 3 elements in it, so we say it has a length of 3. In the example pictured below, the array has 2 axes. The first axis has a length of 2, the second axis has a length of 3.

```
[[1., 0., 0.],
 [0., 1., 2.]]
```

NumPy's array class is called ndarray. It is also known by the alias array. Note that numpy.array is not the same as the Standard Python Library class array.array, which only handles one-dimensional arrays and offers less functionality. The more important attributes of an ndarray object are:

ndarray.ndim

The number of axes (dimensions) of the array.

ndarray.shape

The dimensions of the array. This is a tuple of integers indicating the size of the array in each dimension. For a matrix with *n* rows and *m* columns, shape will be (n,m). The length of the shape tuple is therefore the number of axes, ndim.

ndarray.size

The total number of elements of the array. This is equal to the product of the elements of shape.

ndarray.dtype

An object describing the type of the elements in the array. One can create or specify dtype's using standard Python types. Additionally NumPy provides types of its own. numpy.int32, numpy.int16, and numpy.float64 are some examples.

ndarray.itemsize

The size in bytes of each element of the array. For example, an array of elements of type float64 has itemsize 8 (=64/8), while one of type complex32 has itemsize 4 (=32/8). It is equivalent to ndarray.dtype.itemsize.

ndarray.data

The buffer containing the actual elements of the array. Normally, we won't need to use this attribute because we will access the elements in an array using indexing facilities.

Example

```
>>>import numpy as np
>>> a = np.arange(15).reshape(3, 5)
>>> a
array([[ 0,  1,  2,  3,  4],
       [ 5,  6,  7,  8,  9],
       [10, 11, 12, 13, 14]])
>>> a.shape
```

```
(3, 5)
>>> a.ndim
2
>>> a.dtype.name
'int64'
>>> a.itemsize
8
>>> a.size
15
>>> type(a)
<class 'numpy.ndarray'>
>>> b = np.array([6, 7, 8])
>>> b
array([6, 7, 8])
>>> type(b)
<class 'numpy.ndarray'>
```

Array Creation

There are several ways to create arrays.

For example, you can create an array from a regular Python list or tuple using the array function. The type of the resulting array is deduced from the type of the elements in the sequences.

```
>>>import numpy as np
>>> a = np.array([2, 3, 4])
>>> a
array([2, 3, 4])
>>> a.dtype
dtype('int64')
>>> b = np.array([1.2, 3.5, 5.1])
>>> b.dtype
dtype('float64')
```

array transforms sequences of sequences into two-dimensional arrays, sequences of sequences of sequences into three-dimensional arrays, and so on.

```
>>>b = np.array([(1.5, 2, 3), (4, 5, 6)])
>>> b
array([[1.5, 2. , 3. ],
       [4. , 5. , 6. ]])
```

The type of the array can also be explicitly specified at creation time:

```
>>>c = np.array([[1, 2], [3, 4]], dtype=complex)
>>> c
array([[1.+0.j, 2.+0.j],
       [3.+0.j, 4.+0.j]])
```

Often, the elements of an array are originally unknown, but its size is known. Hence, NumPy offers several functions to create arrays with initial placeholder content. These minimize the necessity of growing arrays, an expensive operation.

The function zeros creates an array full of zeros, the function ones creates an array full of ones, and the function empty creates an array whose initial content is random and depends on the state of the memory. By default, the dtype of the created array is float64, but it can be specified via the key word argument dtype.

```
>>>np.zeros((3, 4))
array([[0., 0., 0., 0.],
       [0., 0., 0., 0.],
       [0., 0., 0., 0.]])
>>> np.ones((2, 3, 4), dtype=np.int16)
array([[[1, 1, 1, 1],
        [1, 1, 1, 1],
        [1, 1, 1, 1]],

       [[1, 1, 1, 1],
        [1, 1, 1, 1],
        [1, 1, 1, 1]]], dtype=int16)
>>> np.empty((2, 3))
array([[3.73603959e-262, 6.02658058e-154, 6.55490914e-260],  # may vary
          [5.30498948e-313, 3.14673309e-307, 1.00000000e+000]])
```

To create sequences of numbers, NumPy provides the arange function which is analogous to the Python built-in range, but returns an array.

```
>>>np.arange(10, 30, 5)
array([10, 15, 20, 25])
>>> np.arange(0, 2, 0.3)  # it accepts float arguments
array([0. , 0.3, 0.6, 0.9, 1.2, 1.5, 1.8])
```

When arange is used with floating point arguments, it is generally not possible to predict the number of elements obtained, due to the finite floating point precision. For this reason, it is usually better to use the function linspace that receives as an argument the number of elements that we want, instead of the step:

```
>>>from numpy import pi
>>> np.linspace(0, 2, 9)                 # 9 numbers from 0 to 2
array([0.  , 0.25, 0.5 , 0.75, 1.  , 1.25, 1.5 , 1.75, 2.  ])
>>> x = np.linspace(0, 2 * pi, 100) # useful to evaluate
                                    #function at lots of points
>>> f = np.sin(x)
```

Printing Arrays

When you print an array, NumPy displays it in a similar way to nested lists, but with the following layout:

- the last axis is printed from left to right,
- the second-to-last is printed from top to bottom,
- the rest are also printed from top to bottom, with each slice separated from the next by an empty line.

One-dimensional arrays are then printed as rows, bidimensionals as matrices and tridimensionals as lists of matrices.

```
>>>a = np.arange(6)                    # 1d array
>>> print(a)
[0 1 2 3 4 5]
>>> b = np.arange(12).reshape(4, 3)     # 2d array
>>> print(b)
[[ 0  1  2]
 [ 3  4  5]
 [ 6  7  8]
 [ 9 10 11]]
>>> c = np.arange(24).reshape(2, 3, 4)  # 3d array
>>> print(c)
[[[ 0  1  2  3]
  [ 4  5  6  7]
  [ 8  9 10 11]]

 [[12 13 14 15]
  [16 17 18 19]
  [20 21 22 23]]]
```

If an array is too large to be printed, NumPy automatically skips the central part of the array and only prints the corners:

```
>>>print(np.arange(10000))
[   0    1    2 ... 9997 9998 9999]
>>>
>>> print(np.arange(10000).reshape(100, 100))
[[   0    1    2 ...   97   98   99]
 [ 100  101  102 ...  197  198  199]
 [ 200  201  202 ...  297  298  299]
 ...
 [9700 9701 9702 ... 9797 9798 9799]
 [9800 9801 9802 ... 9897 9898 9899]
 [9900 9901 9902 ... 9997 9998 9999]]
```

To disable this behaviour and force NumPy to print the entire array, you can change the printing options using set_printoptions.

```
>>> np.set_printoptions(threshold=sys.maxsize)  # sys module should be imported
```

Basic Operations

Arithmetic operators on arrays apply *elementwise*. A new array is created and filled with the result.

```
>>> a = np.array([20, 30, 40, 50])
>>> b = np.arange(4)
>>> b
array([0, 1, 2, 3])
>>> c = a - b
>>> c
array([20, 29, 38, 47])
>>> b**2
array([0, 1, 4, 9])
>>> 10 * np.sin(a)
array([ 9.12945251, -9.88031624,  7.4511316 , -2.62374854])
>>> a < 35
```

```
array([ True,  True, False, False])
```

Unlike in many matrix languages, the product operator * operates elementwise in NumPy arrays. The matrix product can be performed using the @ operator (in python >=3.5) or the dot function or method:

```
>>> A = np.array([[1, 1],
...               [0, 1]])
>>> B = np.array([[2, 0],
...               [3, 4]])
>>> A * B     # elementwise product
array([[2, 0],
       [0, 4]])
>>> A @ B     # matrix product
array([[5, 4],
       [3, 4]])
>>> A.dot(B)  # another matrix product
array([[5, 4],
       [3, 4]])
```

Some operations, such as += and *=, act in place to modify an existing array rather than create a new one.

```
>>> rg = np.random.default_rng(1)  # create instance of default random number generator
>>> a = np.ones((2, 3), dtype=int)
>>> b = rg.random((2, 3))
>>> a *= 3
>>> a
array([[3, 3, 3],
       [3, 3, 3]])
>>> b += a
>>> b
array([[3.51182162, 3.9504637 , 3.14415961],
       [3.94864945, 3.31183145, 3.42332645]])
>>> a += b  # b is not automatically converted to integer type
Traceback (most recent call last):
  ...
numpy.core._exceptions._UFuncOutputCastingError: Cannot cast ufunc 'add' output from dtype('float64') to dtype('int64') with casting rule 'same_kind'
```

When operating with arrays of different types, the type of the resulting array corresponds to the more general or precise one (a behavior known as upcasting).

```
>>> a = np.ones(3, dtype=np.int32)
>>> b = np.linspace(0, pi, 3)
>>> b.dtype.name
'float64'
>>> c = a + b
>>> c
array([1.        , 2.57079633, 4.14159265])
>>> c.dtype.name
'float64'
>>> d = np.exp(c * 1j)
>>> d
array([ 0.54030231+0.84147098j, -0.84147098+0.54030231j,
       -0.54030231-0.84147098j])
>>> d.dtype.name
```

'complex128'

Many unary operations, such as computing the sum of all the elements in the array, are implemented as methods of the ndarray class.

```
>>> a = rg.random((2, 3))
>>> a
array([[0.82770259, 0.40919914, 0.54959369],
       [0.02755911, 0.75351311, 0.53814331]])
>>> a.sum()
3.1057109529998157
>>> a.min()
0.027559113243068367
>>> a.max()
0.8277025938204418
```

Questions

1. Convert a 1-D array into a 2-D array with 3 rows.

Start with:
Assign-1 = np.array([0, 1, 2, 3, 4, 5, 6, 7, 8])
Desired output:
[[0, 1, 2]
[3, 4, 5]
[6, 7, 8]]

2. Replace all odd numbers in the given array with -1

Start with:
Assign-2 = np.array([0, 1, 2, 3, 4, 5, 6, 7, 8, 9])
Desired output:
[0, -1, 2, -1, 4, -1, 6, -1, 8, -1]

3. Find the positions of:

elements in x where its value is more than its corresponding element in y, and elements in x where its value is equals to its corresponding element in y.

Start with these:
x = np.array([21, 64, 86, 22, 74, 55, 81, 79, 90, 89])
y = np.array([21, 7, 3, 45, 10, 29, 55, 4, 37, 18])

Desired output:
(array([1, 2, 4, 5, 6, 7, 8, 9]),) and (array([0]),)

4. **Extract the first four columns of this 2-D array.**
Start with this:
Assign-4= np.arange(100).reshape(5,-1)

Desired output:
[[0 1 2 3]
[20 21 22 23]
[40 41 42 43]
[60 61 62 63]
[80 81 82 83]]

Additional questions

1. **Generate a 1-D array of 10 random integers. Each integer should be a number between 30 and 40 (inclusive).**
Sample of desired output:
[36, 30, 36, 38, 31, 35, 36, 30, 32, 34]

2. Consider the following matrices :
A= ((1, 2, 3), (4, 5, 6), (7, 8, 10)) and B = ((7, 8, 10) ,(4, 5, 6), (1, 2, 3))
Write a python program to perform the following using Numeric Python (numpy).
i) Add and Subtract of the Matrix A and B, print the resultant matrix C for add and E for subtract.
ii) Compute the sum of all elements of Matrix A, sum of each column of Matrix B and sum of each row of Matrix C
iii) Product of two matrices A and B, and print the resultant matrix D
iv) Sort the elements of resultant matrix C and print the resultant Matrix E.
v) Transpose the Matrix E and print the result

WEEK-02: DATA ANALYSIS_AND_VISUALISATION_WITH_PYTHON

Matplotlib

Matplotlib is an amazing visualization library in Python for 2D plots of arrays. Matplotlib is a multi-platform data visualization library built on NumPy arrays and designed to work with the broader SciPy stack. It was introduced by John Hunter in the year 2002. One of the greatest benefits of visualization is that it allows us visual access to huge amounts of data in easily digestible visuals. Matplotlib consists of several plots like line, bar, scatter, histogram etc.

Basic plots in Matplotlib :

Matplotlib comes with a wide variety of plots. Plots helps to understand trends, patterns, and to make correlations. They're typically instruments for reasoning about quantitative information. Some of the sample plots are covered here.

Line plot :

Example 1

```
# importing matplotlib module
from matplotlib import pyplot as plt
# x-axis values
x = [5, 2, 9, 4, 7]
# Y-axis values
y = [10, 5, 8, 4, 2]
# Function to plot
plt.plot(x,y)
# function to show the plot
plt.show()
```

Bar plot:

Example 2

```
# importing matplotlib module
from matplotlib import pyplot as plt
# x-axis values
x = [5, 2, 9, 4, 7]
# Y-axis values
y = [10, 5, 8, 4, 2]
# Function to plot the bar
plt.bar(x,y)
# function to show the plot
plt.show()
```

Hist plot:

A histogram is basically used to represent data provided in a form of some groups.It is accurate method for the graphical representation of numerical data distribution.It is a type of bar plot where X-axis represents the bin ranges while Y-axis gives information about frequency.

Creating a Histogram

To create a histogram the first step is to create bin of the ranges, then distribute the whole range of the values into a series of intervals, and count the values which fall into each of the intervals.Bins are clearly identified as consecutive, non-overlapping intervals of variables.The matplotlib.pyplot.hist() function is used to compute and create histogram of x.

In Matplotlib, we use the hist() function to create histograms. The hist() function will use an array of numbers to create a histogram, the array is sent into the function as an argument. For simplicity we use NumPy to

randomly generate an array with 250 values, where the values will concentrate around 170, and the standard deviation is 10.
You can read from the histogram that there are approximately:
2 people from 140 to 145cm
5 people from 145 to 150cm
15 people from 151 to 156cm
31 people from 157 to 162cm
46 people from 163 to 168cm
53 people from 168 to 173cm
45 people from 173 to 178cm
28 people from 179 to 184cm
21 people from 185 to 190cm
4 people from 190 to 195cm

Example: Say you ask for the height of 250 people, you might end up with a histogram like this:

The hist() function will read the array and produce a histogram:

Simple histogram:

Example 3

```
import matplotlib.pyplot as plt
import numpy as np
x = np.random.normal(170, 10, 250)
print(x)
plt.hist(x)
plt.show()
```

Ref: https://www.w3schools.com/python/matplotlib_histograms.asp

Example 4

```
# importing matplotlib module
from matplotlib import pyplot as plt
# Y-axis values
y = [10, 5, 8, 4, 2]
# Function to plot histogram
plt.hist(y)
# Function to show the plot
plt.show()
```

Ref: https://www.geeksforgeeks.org/plotting-histogram-in-python-using-matplotlib/

Scatter plot:

Example 5

```
# importing matplotlib module
from matplotlib import pyplot as plt
# x-axis values
x = [5, 2, 9, 4, 7]
# Y-axis values
y = [10, 5, 8, 4, 2]
# Function to plot scatter
plt.scatter(x, y)
# function to show the plot
plt.show()
```

SciPy

What is SciPy?

- SciPy is a scientific computation library that uses NumPy underneath.
- SciPy stands for Scientific Python.
- It provides more utility functions for optimization, stats and signal processing.
- Like NumPy, SciPy is open source so we can use it freely.
- SciPy was created by NumPy's creator Travis Olliphant.

Why Use SciPy?

- If SciPy uses NumPy underneath, why can we not just use NumPy?
- SciPy has optimized and added functions that are frequently used in NumPy and Data Science.

Which Language is SciPy Written in?

- SciPy is predominantly written in Python, but a few segments are written in C.

Import SciPy

- Once SciPy is installed, import the SciPy module(s) you want to use in your applications by adding the from scipy import module statement:

  from scipy import constants

constants: SciPy offers a set of mathematical constants, one of them is liter which returns 1 liter as cubic meters.

Unit Categories

- The units are placed under these categories:
- Metric
- Binary
- Mass
- Angle
- Time
- Length
- Pressure
- Volume
- Speed
- Temperature
- Energy
- Power
- Force

Example 6

```
from scipy import constants
print(constants.peta)    #1000000000000000.0
print(constants.tera)    #1000000000000.0
print(constants.giga)    #1000000000.0
print(constants.mega)    #1000000.0
print(constants.kilo)    #1000.0
print(constants.hecto)   #100.0
print(constants.deka)    #10.0
print(constants.deci)    #0.1
print(constants.centi)   #0.01
print(constants.milli)   #0.001
print(constants.micro)   #1e-06
print(constants.nano)    #1e-09
print(constants.pico)    #1e-12
```

Sparse Data

- Sparse data is data that has mostly unused elements (elements that don't carry any information).
- It can be an array like this one: [1, 0, 2, 0, 0, 3, 0, 0, 0, 0, 0, 0]
- **Sparse Data:** is a data set where most of the item values are zero.
- **Dense Array:** is the opposite of a sparse array: most of the values are *not* zero.
- In scientific computing, when we are dealing with partial derivatives in linear algebra we will come across sparse data.

Work With Sparse Data

SciPy has a module, scipy.sparse that provides functions to deal with sparse data.

- There are primarily two types of sparse matrices that we use:
- CSC - Compressed Sparse Column. For efficient arithmetic, fast column slicing.
- CSR - Compressed Sparse Row. For fast row slicing, faster matrix vector products We will use the CSR matrix in this tutorial.

CSR Matrix

We can create CSR matrix by passing an arrray into function scipy.sparse.csr_matrix().

Example 7

Create a CSR matrix from an array:

```
import numpy as np
from scipy.sparse import csr_matrix
arr = np.array([0, 0, 0, 0, 0, 1, 1, 0, 2])
print(csr_matrix(arr))
```

The example above returns:

```
  (0, 5) 1
  (0, 6) 1
  (0, 8) 2
```

From the result we can see that there are 3 items with value.
The 1. item is in row 0 position 5 and has the value 1.
The 2. item is in row 0 position 6 and has the value 1.
The 3. item is in row 0 position 8 and has the value 2.

Viewing stored data (not the zero items) with the data property:

Example 8

```
import numpy as np
from scipy.sparse import csr_matrix
arr = np.array([[0, 0, 0], [0, 0, 1], [1, 0, 2]])
print(csr_matrix(arr).data)
```

Converting from csr to csc with the tocsc() method:

Example 9

```
import numpy as np
from scipy.sparse import csr_matrix
arr = np.array([[0, 0, 0], [0, 0, 1], [1, 0, 2]])
newarr = csr_matrix(arr).tocsc()
print(newarr)
```

Pandas

Pandas is a Python library used for working with data sets. It has functions for analyzing, cleaning, exploring, and manipulating data. The name "Pandas" has a reference to both "Panel Data", and "Python Data Analysis" and was created by Wes McKinney in 2008.

Use of Pandas

Pandas allows us to analyze big data and make conclusions based on statistical theories. Pandas can clean messy data sets, and make them readable and relevant. Relevant data is very important in data science.

Data Science: is a branch of computer science where we study how to store, use and analyze data for deriving information from it.

Pandas gives you answers about the data. Like:

- Is there a correlation between two or more columns?
- What is average value?
- Max value?
- Min value?

Pandas are also able to delete rows that are not relevant, or contains wrong values, like empty or NULL values. This is called *cleaning* the data.

Once Pandas is installed, import it in your applications by adding the import keyword:

import pandas

Example 10

Get your own Python Server

```
import pandas
mydataset = {   'cars': ["BMW", "Volvo", "Ford"],
  'passings': [3, 7, 2] }
myvar = pandas.DataFrame(mydataset)
print(myvar)
```

Create an alias with the as keyword while importing:

import pandas as pd

Now the Pandas package can be referred to as pd instead of pandas.

Example 11

```
import pandas as pd
mydataset = {
  'cars': ["BMW", "Volvo", "Ford"],
  'passings': [3, 7, 2]
}
myvar = pd.DataFrame(mydataset)
print(myvar)
```

Indices in a pandas series:

- A pandas series is similar to a list, but differs in the fact that a series associates a label with each element. This makes it look like a dictionary.
- If an index is not explicitly provided by the user, pandas creates a RangeIndex ranging from 0 to N-1.
- Each series object also has a data type.

Example 12

```
import panadad as pd
new_series= pd.series([5,6,7,8,9,10])
print(new_series)
```

- As you may suspect by this point, a series has ways to extract all of the values in the series, as well as individual elements by index.

Example 13

```
import panadad as pd
new_series= pd.series([5,6,7,8,9,10])
print(new_series.values)
print('_____')
print(new_series[4])
```

- You can also provide an index manually.

Example 14

```
import panadad as pd
new_series= pd.series([5,6,7,8,9,10], indec=['a','b','c','d','e','f')
print(new_series.values)
print('_____________')
print(new_series['f'])
```

- It is easy to retrieve several elements of a series by their indices or make group assignments.

Example 15

```
import panadad as pd
new_series= pd.series([5,6,7,8,9,10], indec=['a','b','c','d','e','f')
print(new_series)
print('_____________')
print(new_series['a','b','f'])=0
print(new_series)
```

- Filtering and maths operations are easy with Pandas as well.

Example 16

```
import panadad as pd
new_series= pd.series([5,6,7,8,9,10], indec=['a','b','c','d','e','f')
new_series2= new_series[new_series>0]  # check with 7 instead of 0
print(new_series2)
print('_____________')
new_series2= new_series[new_series>0]  * 2
print(new_series2)
```

DataFrame:

- A Pandas DataFrame is a 2 dimensional data structure, like a 2 dimensional array, or a table with rows and columns.
- Simplistically, a data frame is a table, with rows and columns.
- Each column in a data frame is a series object.
- Rows consist of elements inside series.

Case ID	*Variable one*	*Variable two*	*Variable 3*
1	*123*	*ABC*	*10*
2	*456*	*DEF*	*20*
3	*789*	*XYZ*	*30*

- Pandas data frames can be constructed using Python dictionaries.

Example 17

Get your own Python Server

Create a simple Pandas DataFrame:

```
import pandas as pd
data = {
  "calories": [420, 380, 390],
  "duration": [50, 40, 45]
}
#load data into a DataFrame object:
df = pd.DataFrame(data)
print(df)
```

Locate Row

As you can see from the result above, the DataFrame is like a table with rows and columns. Pandas use the loc attribute to return one or more specified row(s)

Example 18

Return row 0:

```
#refer to the row index:
print(df.loc[0])
```

Named Indexes

With the index argument, you can name your own indexes.

Example 19

Add a list of names to give each row a name:

```
import pandas as pd
data = {
  "calories": [420, 380, 390],
  "duration": [50, 40, 45]
}
df = pd.DataFrame(data, index = ["day1", "day2", "day3"])
print(df)
```

- You can also create a data frame from a list.

Example 20

```
import pandas as pd
list2=[[0,1,2],[3,4,5],[6,7,8]]
df= pd.DataFrame(list2)
print(df)
df.columns=['V1','V2','V3']
print(df)
```

Example 21

```
import pandas as pd
df=pd.DataFrame({
    'Country': ['Kazakhstan','Russia','Belarus','Ukraine'],
    'Population': [17.04,143.5,9.5,45.5]
    'Square':[2724902, 17125191,207600,603628]
   })
print(df)
```

- You can ascertain the type of a column with the type() function.

```
print(type(df['Country']))
```

- A Pandas data frame object as two indices; a column index and row index.
- Again, if you do not provide one, Pandas will create a RangeIndex from 0 to N-1.

Example 22

```
import pandas as pd
df=pd.DataFrame({
    'Country': ['Kazakhstan','Russia','Belarus','Ukraine'],
    'Population': [17.04,143.5,9.5,45.5]
    'Square':[2724902, 17125191,207600,603628]
   })
print(df.columns)
print('_____')
print(df.index)
```

- There are numerous ways to provide row indices explicitly.
- For example, you could provide an index when creating a data frame:

Example 23

```
import pandas as pd
df=pd.DataFrame({
    'Country': ['Kazakhstan','Russia','Belarus','Ukraine'],
    'Population': [17.04,143.5,9.5,45.5]
    'Square':[2724902, 17125191,207600,603628]
   }, index = ['KZ','RU','BY','UA'])
print(df)
```

- or do it during runtime.
- Here, I also named the index ‘country code’.

Example 24

```
import pandas as pd
df=pd.DataFrame({
    ‘Country’: [‘Kazakhstan’,’Russia’,’Belarus’,’Ukraine’],
    ‘Population’: [17.04,143.5,9.5,45.5]
    ‘Square’:[2724902, 17125191,207600,603628]
   })
print(df)
print(‘_____’)
df.index = =[‘KZ’,’RU’,’BY’,’UA’]
df.index.name = ‘Country Code’
print(df)
```

- Row access using index can be performed in several ways.
- First, you could use .loc() and provide an index label.

Example 25

```
print(df.loc[‘KZ’]
```

- Second, you could use .iloc() and provide an index number

```
print(df.iloc[0]
```

- A selection of particular rows and columns can be selected this way.

```
print(df.loc[[‘KZ’,’RU’],’population’])
```

- You can feed .loc() two arguments, index list and column list, slicing operation is supported as well:

```
print(df.loc[[‘KZ’,’BY’, :])
```

Filtering

- Filtering is performed using so-called Boolean arrays.

Example 26

```
print(df[df.population > 10][[‘Country’,’Square’]])
```

Deleting columns

You can delete a column using the drop() function.

```
print(df)
df = df.drop([‘Population’], axis = ‘columns’)
print(df)
```

Load Files Into a DataFrame

If your data sets are stored in a file, Pandas can load them into a DataFrame.

Example 27

Load a comma separated file (CSV file) into a DataFrame:

```
import pandas as pd
df = pd.read_csv('data.csv')
print(df)
```

Read CSV Files

A simple way to store big data sets is to use CSV files (comma separated files). CSV files contains plain text and is a well know format that can be read by everyone including Pandas. In our examples we will be using a CSV file called 'data.csv'.

Example 28

Get your own Python Server

Load the CSV into a DataFrame:

```
import pandas as pd
df = pd.read_csv('data.csv')
print(df.to_string())  # use to_string() to print the entire DataFrame.
print(df)   #Print the DataFrame without the to_string() method
```

Data Cleaning

Data cleaning means fixing bad data in your data set.

Bad data could be:

- Empty cells
- Data in wrong format
- Wrong data
- Duplicates

Plotting

Pandas uses the plot() method to create diagrams. We can use Pyplot, a submodule of the Matplotlib library to visualize the diagram on the screen.

Example 29

Get your own Python Server

Import pyplot from Matplotlib and visualize our DataFrame:

```
import pandas as pd
import matplotlib.pyplot as plt
df = pd.read_csv('data.csv')
df.plot()
plt.show()
```

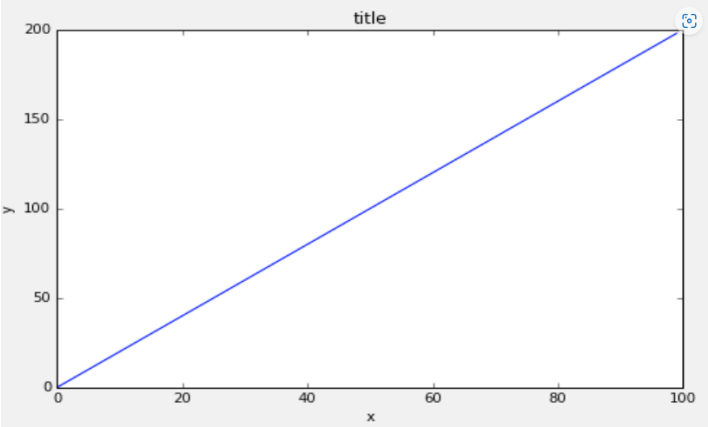
Questions

1. Follow along with these steps:

a) Create a figure object called fig using plt.figure()

b) Use add_axes to add an axis to the figure canvas at [0,0,1,1]. Call this new axis ax.

c) Plot (x,y) on that axes and set the labels and titles to match the plot below:



2. Create a figure object and put two axes on it, ax1 and ax2. Located at [0,0,1,1] and [0.2,0.5,.2,.2] respectively. Now plot (x,y) on both axes. And call your figure object to show it.

3. Use the company sales dataset csv file, read Total profit of all months and show it using a line plot Total profit data provided for each month. Generated line plot must include the following properties: –

a. X label name = Month Number
b. Y label name = Total profit

4. Use the company sales dataset csv file, get total profit of all months and show line plot with the following Style properties. Generated line plot must include following Style properties: –

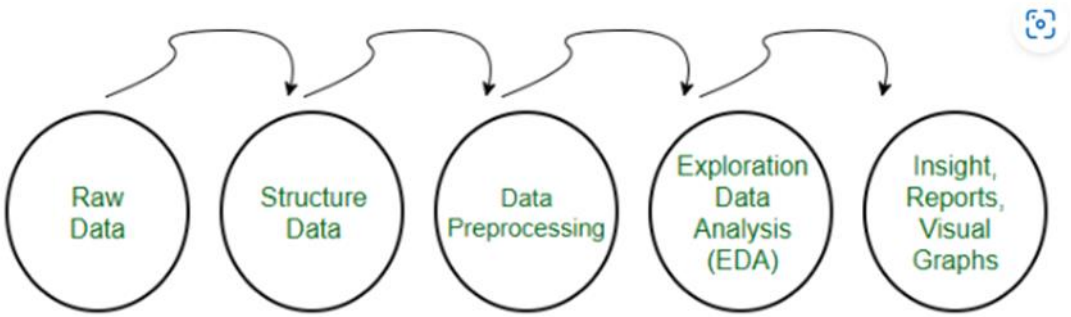
a. Line Style dotted and Line-color should be red
b. Show legend at the lower right location.
c. X label name = Month Number
d. Y label name = Sold units number
e. Add a circle marker.
f. Line marker color as read
g. Line width should be 3

Additional Questions

1. Use the company sales dataset csv file, read all product sales data and show it using a multiline plot. Display the number of units sold per month or each product using multiline plots. (i.e., Separate Plotline for each product).
2. Use the company sales dataset csv file, calculate total sale data for last year for each product and show it using a Pie chart.
   Note: In Pie chart display Number of units sold per year for each product in percentage.

WEEK -03 : DATA PREPROCESSING AND REGRESSION

Data pre-processing is a basic requirement of any good machine learning model. Pre-processing the data implies using the data which is easily readable by the machine learning model. In this week, we are going to discuss the basics of data pre-processing and how to make the data suitable for machine learning models.



Data Preprocessing

What is data pre-processing?

Data pre-processing is the process of preparing the raw data and making it suitable for machine learning models. Data pre-processing includes data cleaning for making the data ready to be given to machine learning model.

After data cleaning, data pre-processing requires the data to be transformed into a format that is understandable to the machine learning model.

Why is data pre-processing required?

Data pre-processing is mainly required for the following:

- **Accurate data:**For making the data readable for machine learning model, it needs to be accurate with no missing value, redundant or duplicate values.
- **Trusted data:**The updated data should be as accurate or trusted as possible.
- **Understandable data:**The data updated needs to be interpreted correctly.

All in all, data pre-processing is important for the machine learning model to learn from such data which is correct in order to lead the model to the right predictions/outcomes.

Examples of data preprocessing for different data set types with Python

Since data comes in various formats, let us discuss how different data types can be converted into a format that the ML model can read accurately. Let us see how to feed correct features from datasets with:

- Missing values
- Outliers
- Overfitting
- Data with no numerical values
- Different date formats

▪ *Missing values*

Missing values are a common problem while dealing with data! The values can be missed because of various reasons such as human errors, mechanical errors, etc.

Data cleansing is an important step before you even begin the algorithmic trading process, which begins with historical data analysis for making the prediction model as accurate as possible.

Based on this prediction model you create the trading strategy. Hence, leaving missed values in the data set can wreak havoc by giving faulty predictive results that can lead to erroneous strategy creation and further the results can not be great to state the obvious.

There are three techniques to solve the missing values' problem in order to find out the most accurate features, and they are:

Dropping
Numerical imputation
Categorical imputation

Dropping

Dropping is the most common method to take care of the missed values. Those rows in the data set or the entire columns with missed values are dropped in order to avoid errors to occur in data nalysis.

There are some machines that are programmed to automatically drop the rows or columns that include missed values resulting in a reduced training size. Hence, the dropping can lead to a decrease in the model performance.

A simple solution for the problem of a decreased training size due to the dropping of values is to use imputation. We will discuss the interesting imputation methods further. In case of dropping, you can define a threshold to the machine.

For instance, the threshold can be anything. It can be 50%, 60% or 70% of the data. Let us take 60% in our example, which means that 60% of data with missing out values will be accepted by the model/algorithm as the training data set, but the features with more than 60% missing values will be dropped.

For dropping the values, following Python codes are used:

```
#Dropping columns in the data higher than 60% threshold
data = data[data.columns[data.isnull().mean() < threshold]]
#Dropping rows in the data higher than 60% threshold
data = data.loc[data.isnull().mean(axis=1) < threshold]
```

By using the above Python codes, the missed values will be dropped and the machine learning model will learn on the rest of the data.

Numerical imputation

The word imputation implies replacing the missing values with such a value that makes sense. And, numerical imputation is done in the data with numbers.

For instance, if there is a tabular data set with the number of stocks, commodities and derivatives traded in a month as the columns, it is better to replace the missed value with a "0" than leaving them as it is.

With numerical imputation, the data size is preserved and hence, predictive models like linear regression can work better to predict in the most accurate manner.

A linear regression model can not work with missing values in the data set since it is biased toward the missed values and considers them "good estimates". Also, the missed values can be replaced with the median of the columns since median values are not sensitive to outliers unlike averages of columns.

Let us see the Python codes for numerical imputation, which are as follows:

```
#For filling all the missed values as 0
data = data.fillna(0)
#For replacing missed values with median of columns
data = data.fillna(data.median())
```

Categorical imputation

This technique of imputation is nothing but replacing the missed values in the data with the one which occurs the maximum number of times in the column. But, in case there is no such value that occurs frequently or dominates the other values, then it is best to fill the same as "NAN".

The following Python code can be used here:

```
#Categorical imputation
data['column_name'].fillna(data['column_name'].value_counts().idxmax(), inplace=True)
```

- *Outliers*

An outlier differs significantly from other values and is too distanced from the mean of the values. Such values that are considered outliers are usually due to some systematic errors or flaws.

Let us see the following Python codes for identifying and removing outliers with standard deviation:

```
#For identifying the outliers with the standard deviation method
outliers = [x for x in data if x < lower or x > upper]
print('Identified outliers: %d' % len(outliers))
#Remove outliers
outliers_removed = [x for x in data if x >= lower and x <= upper]
print('Non-outlier observations: %d' % len(outliers_removed))
```

In the codes above, “lower” and “upper” signify the upper and lower limit in the data set.

- *Overfitting*

In both machine learning and statistics, overfitting occurs when the model fits the data too well or simply put when the model is too complex.

Overfitting model learns the detail and noise in the training data to such an extent that it negatively impacts the performance of the model on new data/test data.

The overfitting problem can be solved by decreasing the number of features/inputs or by increasing the number of training examples to make the machine learning algorithms more generalised.

The most common solution is regularisation in an overfitting case. Binning is the technique that helps with the regularisation of the data which also makes you lose some data every time you regularise it.

For instance, in the case of numerical binning, the data can be as follows:

Stock value	Bin
100-250	Lowest
251-400	Mid
401-500	High

Here is the Python code for binning:

```
data['bin'] = pd.cut(data['value'], bins=[100,250,400,500], labels=["Lowest", "Mid", "High"])
```

Your output should look something like this:

	Value	Bin
0	102	Low
1	300	Mid
2	107	Low
3	470	High

- *Data with no numerical values*

In the case of the data set with no numerical values, it becomes impossible for the machine learning model to learn the information.

The machine learning model can only handle numerical values and thus, it is best to spread the values in the columns with assigned binary numbers “0” or “1”. This technique is known as one-hot encoding.

In this type of technique, the grouped columns already exist. For instance, below I have mentioned a grouped column:

Infected	Covid variants
2	Delta
4	Lambda
5	Omicron
6	Lambda
4	Delta
3	Omicron
5	Omicron
4	Lambda
2	Delta

Now, the above-grouped data can be encoded with the binary numbers ”0” and “1” with one hot encoding technique. This technique subtly converts the categorical data into a numerical format in the following manner:

Infected	Delta	Lambda	Omicron
2	1	0	0
4	0	1	0
5	0	0	1
6	0	1	0
4	1	0	0
3	0	0	1
5	0	0	1
4	0	1	0
2	1	0	0

Hence, it results in better handling of grouped data by converting the same into encoded data for the machine learning model to grasp the encoded (which is numerical) information quickly.

Problem with the approach

Going further, in case there are more than three categories in a data set that is to be used for feeding the machine learning model, the one-hot encoding technique will create as many columns. Let us say, there are 2000 categories, then this technique will create 2000 columns and it will be a lot of information to feed to the model.

Solution:
To solve this problem, while using this technique, we can apply the target encoding technique which implies calculating the "mean" of each predictor category and using the same mean for all other rows with the same category under the predictor column. This will convert the categorical column into the numeric column and that is our main aim.
Let us understand this with the same example as above but this time we will use the "mean" of the values under the same category in all the rows. Let us see how.
In Python, we can use the following code:

```
#Convert data into numerical values with mean
Infected = [2, 4, 5, 6, 4, 3]
Predictor = ['Delta', 'Lambda', 'Omicron', 'Lambda', 'Delta', 'Omicron']
Infected_df = pd.DataFrame(data={'Infected':Infected, 'Predictor':Predictor})
means = Infected_df.groupby('Predictor')['Infected'].mean()
Infected_df['Predictor_encoded'] = Infected_df['predictor'].map(means)
Infected_df
```

Output:

Infected	Predictor	Predictor_encoded
2	Delta	3
4	Lambda	5
5	Omicron	4
6	Lambda	5
4	Delta	3
3	Omicron	4

In the output above, the Predictor column depicts the Covid variants and the Predictor_encoded column depicts the "mean" of the same category of Covid variants which makes 2+4/2 = 3 as the mean value for Delta, 4+6/2 = 5 as the mean value for Lambda and so on.
Hence, the machine learning model will be able to feed the main feature (converted to a number) for each predictor category for the future.

- *Different date formats*

With the different date formats such as "25-12-2021", "25th December 2021" etc. the machine learning model needs to be equipped with each of them. Or else, it is difficult for the machine learning model to understand all the formats.

With such a data set, you can preprocess or decompose the data by mentioning three different columns for the parts of the date, such as Year, Month and Day.
In Python, the preprocessing of the data with different columns for the date will look like this:

```
#Convert to datetime object
df['Date'] = pd.to_datetime(df['Date'])
#Decomposition
df['Year'] = df['Date'].dt.year
df['Month'] = df['Date'].dt.month
df['Day'] = df['Date'].dt.day
df[['Year','Month','Day']].head()
```

Output:

Year	Month	Day
2019	1	5
2019	3	8
2019	3	3
2019	1	27
2019	2	8

In the output above, the data set is in date format which is numerical. And because of decomposing the date in different parts such as Year, Month and Day, the machine learning model will be able to learn the date format.

REGRESSION

Regression is a supervised learning technique that supports finding the correlation among variables.
A regression problem is when the output variable is a real or continuous value.
What is a Regression?

In Regression, we plot a graph between the variables which best fit the given data points. The machine learning model can deliver predictions regarding the data. In naïve words, **"Regression shows a line or curve that passes through all the data points on a target-predictor graph in such a way that the vertical distance between the data points and the regression line is minimum."** *It is used principally for prediction, forecasting, time series modeling, and determining the causal-effect relationship between variables.*

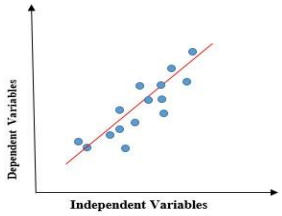
TYPES OF REGRESSION MODELS

1. Linear Regression
2. Polynomial Regression
3. Logistics Regression

1. Linear Regression:

Linear regression is a quiet and simple statistical regression method used for predictive analysis and shows the relationship between the continuous variables. Linear regression shows the linear relationship between the independent variable (X-axis) and the dependent variable (Y-axis), consequently called linear regression. If there is a single input variable (x), such linear regression is called **simple linear regression**. And if there

is more than one input variable, such linear regression is called **multiple linear regression.** The linear regression model gives a sloped straight line describing the relationship within the variables.



The above graph presents the linear relationship between the dependent variable and independent variables. When the value of x (**independent variable**) increases, the value of y (**dependent variable**) is likewise increasing. The red line is referred to as the best fit straight line. Based on the given data points, we try to plot a line that models the points the best.

To calculate best-fit line linear regression uses a traditional slope intercept form.

$$y = mx+b \quad \Longrightarrow \quad y = a_0+a_1x$$

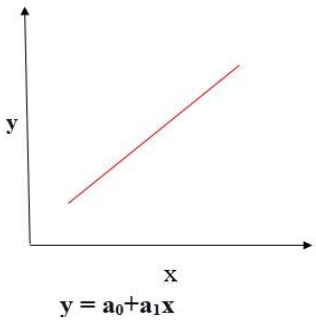
y= Dependent Variable.
x= Independent Variable.
a0= intercept of the line.
a1 = Linear regression coefficient.
Need of a Linear regression:

Linear regression estimates the relationship between a dependent variable and an independent variable. Let's say we want to estimate the salary of an employee based on year of experience. You have the recent company data, which indicates that the relationship between experience and salary. Here year of experience is an independent variable, and the salary of an employee is a dependent variable, as the salary of an employee is dependent on the experience of an employee. Using this insight, we can predict the future salary of the employee based on current & past information.

A regression line can be a Positive Linear Relationship or a Negative Linear Relationship.

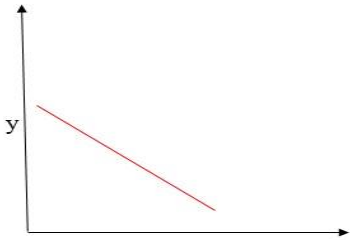
Positive Linear Relationship:

If the dependent variable expands on the Y-axis and the independent variable progress on X-axis, then such a relationship is termed a Positive linear relationship.



Negative Linear Relationship:

If the dependent variable decreases on the Y-axis and the independent variable increases on the X-axis, such a relationship is called a negative linear relationship.



$y = -a_0 + a_1 x$

The goal of the linear regression algorithm is to get the best values for a0 and a1 to find the best fit line. The best fit line should have the least error means the error between predicted values and actual values should be minimized.

Cost function

The cost function helps to figure out the best possible values for a0 and a1, which provides the best fit line for the data points.

Cost function optimizes the regression coefficients or weights and measures how a linear regression model is performing. The cost function is used to find the accuracy of the **mapping function** that maps the input variable to the output variable. This mapping function is also known as **the Hypothesis function**.

In Linear Regression, **Mean Squared Error (MSE)** cost function is used, which is the average of squared error that occurred between the predicted values and actual values.

By simple linear equation y=mx+b we can calculate MSE as:

Let's y = actual values, y_i = predicted values

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(y_i - (mx_i + b))^2$$

Using the MSE function, we will change the values of a0 and a1 such that the MSE value settles at the minima. Model parameters ***xi, b (a_0,a_1)*** can be manipulated to minimize the cost function. These parameters can be determined using the gradient descent method so that the cost function value is minimum.

SCIKIT LEARN

- It is mainly used in machine learning
- It has lot of statistics related tools
- It is open source.
- By using the Scikit library the efficiency will improve tremendously as it is quite accurate.
- It is very useful in algorithms which are very famous in machine learning like K-mean, K-nearest, clustering etc.
- It is available to everybody so any programmer if he or she feels like utilizing it then can use it.
- Scikit requires Numpy

Features of Scikit learn are as follows:

- Clustering: Scikit can be applied in clustering algorithm, in clustering the grouping is done on the basis of similarities like eg: age, color etc.
- Cross validation
    - *Feature selection*

Example:

```
from sklearn.datasets import load_iris
iris = load_iris()
A= iris.data
y = iris.target
feature_names = iris.feature_names
target_names = iris.target_names
print("Feature names:", feature_names)
```

```
print("Target names:", target_names)
print("\nFirst 10 rows of A:\n", A[:10])
```

How to split a Dataset into Train and Test Sets using Python

Scikit-learn alias **sklearn** is the most useful and robust library for machine learning in Python. The **scikit-learn library** provides us with the model_selection module in which we have the splitter function train_test_split().

Syntax:

train_test_split(*arrays, test_size=None, train_size=None, random_state=None, shuffle=True, stratify=None)

Parameters:

1. *arrays: inputs such as lists, arrays, data frames, or matrices
2. test_size: this is a float value whose value ranges between 0.0 and 1.0. it represents the proportion of our test size. its default value is none.
3. train_size: this is a float value whose value ranges between 0.0 and 1.0. it represents the proportion of our train size. its default value is none.
4. random_state: this parameter is used to control the shuffling applied to the data before applying the split. it acts as a seed.
5. shuffle: This parameter is used to shuffle the data before splitting. Its default value is true.
6. stratify: This parameter is used to split the data in a stratified fashion.

Example

```
# read the dataset
# import modules
import pandas as pd
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
df = pd.read_csv('/home/student/Downloads/Real-estate.csv')
 # get the locations
X = df.iloc[:, :-1]
y = df.iloc[:, -1]
# split the dataset
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.05, random_state=0)
```

In the above example, **We import the pandas package and sklearn package**. after that to import the CSV file we use the read_csv() method. The variable df now contains the data frame. in the example “house price” is the column we’ve to predict so we take that column as y and the rest of the columns as our X variable. test_size = 0.05 specifies only 5% of the whole data is taken as our test set, and 95% as our train set. The random state helps us get the same random split each time.

Simple Linear Regression With scikit-learn

You’ll start with the simplest case, which is simple linear regression. There are five basic steps when you’re implementing linear regression:

1. Import the packages and classes that you need.
2. Provide data to work with, and eventually do appropriate transformations.
3. Create a regression model and fit it with existing data.
4. Check the results of model fitting to know whether the model is satisfactory.
5. Apply the model for predictions.

These steps are more or less general for most of the regression approaches and implementations. Throughout the rest of the tutorial, you'll learn how to do these steps for several different scenarios.

Step 1: Import packages and classes

The first step is to import the package numpy and the class LinearRegression from sklearn.linear_model:

```
>>> import numpy as np
>>> from sklearn.linear_model import LinearRegression
```

Now, you have all the functionalities that you need to implement linear regression.

The fundamental data type of NumPy is the array type called numpy.ndarray. The rest of this tutorial uses the term array to refer to instances of the type numpy.ndarray.

You'll use the class sklearn.linear_model.LinearRegression to perform linear and polynomial regression and make predictions accordingly.

Step 2: Provide data

The second step is defining data to work with. The inputs (regressors, x) and output (response, y) should be arrays or similar objects. This is the simplest way of providing data for regression:

:

```
>>> x = np.array([5, 15, 25, 35, 45, 55]).reshape((-1, 1))
>>> y = np.array([5, 20, 14, 32, 22, 38])
```

Now, you have two arrays: the input, x, and the output, y. You should call .reshape() on x because this array must be **two-dimensional**, or more precisely, it must have **one column** and **as many rows as necessary**. That's exactly what the argument (-1, 1) of .reshape() specifies.

This is how x and y look now:

```
>>> x
array([[ 5],
       [15],
       [25],
       [35],
       [45],
       [55]])

>>> y
array([ 5, 20, 14, 32, 22, 38])
```

As you can see, x has two dimensions, and x.shape is (6, 1), while y has a single dimension, and y.shape is (6,).

Step 3: Create a model and fit it

The next step is to create a linear regression model and fit it using the existing data.

Create an instance of the class LinearRegression, which will represent the regression model:

```
>>> model = LinearRegression()he
```

This statement creates the variable model as an instance of LinearRegression. You can provide several optional parameters to LinearRegression:

- fit_intercept is a Boolean that, if True, decides to calculate the intercept b_0 or, if False, considers it equal to zero. It defaults to True.
- normalize is a Boolean that, if True, decides to normalize the input variables. It defaults to False, in which case it doesn't normalize the input variables.
- copy_X is a Boolean that decides whether to copy (True) or overwrite the input variables (False). It's True by default.
- n_jobs is either an integer or None. It represents the number of jobs used in parallel computation. It defaults to None, which usually means one job. -1 means to use all available processors.

Your model as defined above uses the default values of all parameters.

It's time to start using the model. First, you need to call .fit() on model:

```
>>> model.fit(x, y)
LinearRegression()
```

With .fit(), you calculate the optimal values of the weights b_0 and b_1, using the existing input and output, x and y, as the arguments. In other words, .fit() **fits the model**. It returns self, which is the variable model itself. That's why you can replace the last two statements with this one:

```
>>> model = LinearRegression().fit(x, y)
```

This statement does the same thing as the previous two. It's just shorter.

Step 4: Get results

Once you have your model fitted, you can get the results to check whether the model works satisfactorily and to interpret it.

You can obtain the coefficient of determination, R^2, with .score() called on model:

```
>>> r_sq = model.score(x, y)
>>> print(f"coefficient of determination: {r_sq}")
coefficient of determination: 0.7158756137479542
```

When you're applying .score(), the arguments are also the predictor x and response y, and the return value is R^2.

The attributes of model are .intercept_, which represents the coefficient b_0, and .coef_, which represents b_1:

```
>>> print(f"intercept: {model.intercept_}")
intercept: 5.633333333333329

>>> print(f"slope: {model.coef_}")
slope: [0.54]
```

The code above illustrates how to get b_0 and b_1. You can notice that .intercept_ is a scalar, while .coef_ is an array.

Note: In scikit-learn, by convention, a trailing underscore indicates that an attribute is estimated. In this example, .intercept_ and .coef_ are estimated values.

The value of b_0 is approximately 5.63. This illustrates that your model predicts the response 5.63 when x is zero. The value $b_1 = 0.54$ means that the predicted response rises by 0.54 when x is increased by one.

You'll notice that you can provide y as a two-dimensional array as well. In this case, you'll get a similar result. This is how it might look:

```
>>> new_model = LinearRegression().fit(x, y.reshape((-1, 1)))
>>> print(f"intercept: {new_model.intercept_}")
intercept: [5.63333333]

>>> print(f"slope: {new_model.coef_}")
slope: [[0.54]]
```

As you can see, this example is very similar to the previous one, but in this case, .intercept_ is a one-dimensional array with the single element b_0, and .coef_ is a two-dimensional array with the single element b_1.

Step 5: Predict response

Once you have a satisfactory model, then you can use it for predictions with either existing or new data. To obtain the predicted response, use .predict():

```
>>> y_pred = model.predict(x)
>>> print(f"predicted response:\n{y_pred}")
predicted response:
[ 8.33333333 13.73333333 19.13333333 24.53333333 29.93333333 35.33333333]
```

When applying .predict(), you pass the regressor as the argument and get the corresponding predicted response. This is a nearly identical way to predict the response:

```
>>> y_pred = model.intercept_ + model.coef_ * x
>>> print(f"predicted response:\n{y_pred}")
predicted response:
[[ 8.33333333]
 [13.73333333]
 [19.13333333]
 [24.53333333]
 [29.93333333]
 [35.33333333]]
```

In this case, you multiply each element of x with model.coef_ and add model.intercept_ to the product.

The output here differs from the previous example only in dimensions. The predicted response is now a two-dimensional array, while in the previous case, it had one dimension.

If you reduce the number of dimensions of x to one, then these two approaches will yield the same result. You can do this by replacing x with x.reshape(-1), x.flatten(), or x.ravel() when multiplying it with model.coef_.

In practice, regression models are often applied for forecasts. This means that you can use fitted models to calculate the outputs based on new inputs:

```
>>> x_new = np.arange(5).reshape((-1, 1))
>>> x_new
array([[0],
       [1],
       [2],
       [3],
```

[4]])

```
>>> y_new = model.predict(x_new)
>>> y_new
array([5.63333333, 6.17333333, 6.71333333, 7.25333333, 7.79333333])
```

Here .predict() is applied to the new regressor x_new and yields the response y_new. This example conveniently uses arange() from numpy to generate an array with the elements from 0, inclusive, up to but excluding 5—that is, 0, 1, 2, 3, and 4.

Ref: https://realpython.com/linear-regression-in-python/

Question

1. Consider the hepatitis/ pima-indians-diabetes csv file, perform the following date pre-processing.
    1. Load data in Pandas.
    2. Drop columns that aren't useful.
    3. Drop rows with missing values.
    4. Create dummy variables.
    5. Take care of missing data.
    6. Convert the data frame to NumPy.
    7. Divide the data set into training data and test data.

2. a. Construct a CSV file with the following attributes:
Study time in hours of ML lab course (x)
Score out of 10 (y)
The dataset should contain 10 rows.
b. Create a regression model and display the following:
Coefficients: B0 (intercept) and B1 (slope)
RMSE (Root Mean Square Error)
Predicted responses
c. Create a scatter plot of the data points in red color and plot the graph of x vs. predicted y in blue color.
d. Implement the model using two methods:
Pedhazur formula (intuitive)
Calculus method (partial derivatives, refer to class notes)
e. Compare the coefficients obtained using both methods and compare them with the analytical solution.
f. Test your model to predict the score obtained when the study time of a student is 10 hours.
Note: Do not use scikit-learn.

Additional Question

1. a. Consider the hepatitis/diabetes CSV file. Create a regression model and display the following:

- Coefficients: B0 (intercept) and B1 (slope)
- RMSE (Root Mean Square Error)
- Predicted responses

b. Create a scatter plot of the data points in red color and plot the graph of x vs. predicted y in blue color.
c. Implement the model using two methods:
    1. Pedhazur formula (intuitive)
    2. Calculus method (partial derivatives, refer to class notes)

d. Compare the coefficients obtained using both methods. For a given data point, check the predicted y value.
Note: Do not use scikit-learn.

WEEK-04: Polynomial and Multiple Regression

Machine Learning Model evaluation metrics

The various ways to check the performance of our machine learning or deep learning model and why to use one in place of the other. We will discuss terms like:

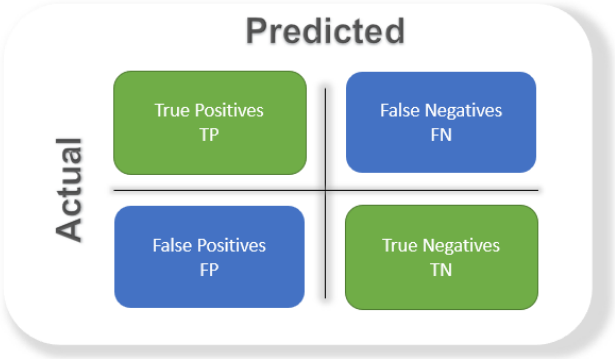
- Confusion matrix
- Accuracy
- Precision
- Recall
- Specificity
- F1 score
- Precision-Recall or PR curve
- **ROC** (**R**eceiver **O**perating **C**haracteristics) curve
- PR vs ROC curve.

For simplicity, we will mostly discuss things in terms of a binary classification problem where let's say we'll have to find if an image is of a cat or a dog. Or a patient is having cancer (positive) or is found healthy (negative). Some common terms to be clear with are:

True positives (TP): Predicted positive and are actually positive.
False positives (FP): Predicted positive and are actually negative.
True negatives (TN): Predicted negative and are actually negative.
False negatives (FN): Predicted negative and are actually positive.

Confusion matrix

It's just a representation of the above parameters in a matrix format. Better visualization is always good :)



Accuracy

The most commonly used metric to judge a model and is actually not a clear indicator of the performance. The worse happens when classes are imbalanced.

$$\frac{TP + TN}{TP + FP + TN + FN}$$

Take for example a cancer detection model. The chances of actually having cancer are very low. Let's say out of 100, 90 of the patients don't have cancer and the remaining 10 actually have it. We don't want to miss on a patient who is having cancer but goes undetected (false negative). Detecting everyone as not having cancer gives an accuracy of 90% straight. The model did nothing here but just gave cancer free for all the 100 predictions.

We surely need better alternatives.

Precision

Percentage of positive instances out of the ***total predicted positive*** instances. Here denominator is the model prediction done as positive from the whole given dataset. Take it as to find out '*how much the model is right when it says it is right'.*

$$\frac{TP}{TP + FP}$$

Recall/Sensitivity/True Positive Rate

Percentage of positive instances out of the ***total actual positive*** instances. Therefore denominator (*TP + FN)* here is the *actual* number of positive instances present in the dataset. Take it as to find out '*how much extra right ones, the model missed when it showed the right ones'.*

$$\frac{TP}{TP + FN}$$

Specificity

Percentage of negative instances out of the ***total actual negative*** instances. Therefore denominator (*TN + FP)* here is the *actual* number of negative instances present in the dataset. It is similar to recall but the shift is on the negative instances. *Like finding out how many healthy patients were not having cancer and were told they don't have cancer*. Kind of a measure to see how separate the classes are.

$$\frac{TN}{TN + FP}$$

F1 score

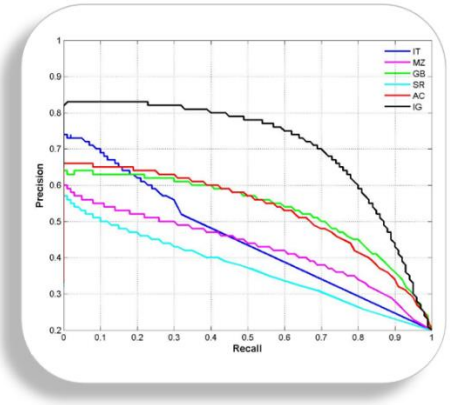
It is the harmonic mean of precision and recall. This takes the contribution of both, so higher the F1 score, the better. See that due to the product in the numerator if one goes low, the final F1 score goes down significantly. So a model does well in F1 score if the positive predicted are actually positives (precision) and doesn't miss out on positives and predicts them negative (recall).

$$\frac{2}{\frac{1}{precision} + \frac{1}{recall}} = \frac{2 * precision * recall}{precision + recall}$$

One drawback is that both precision and recall are given equal importance due to which according to our application we may need one higher than the other and F1 score may not be the exact metric for it. Therefore either weighted-F1 score or seeing the PR or ROC curve can help.

PR curve

It is the curve between precision and recall for various threshold values. In the figure below we have 6 predictors showing their respective precision-recall curve for various threshold values. The top right part of the graph is the ideal space where we get high precision and recall. Based on our application we can choose the predictor and the threshold value. PR AUC is just the area under the curve. The higher its numerical value the better.

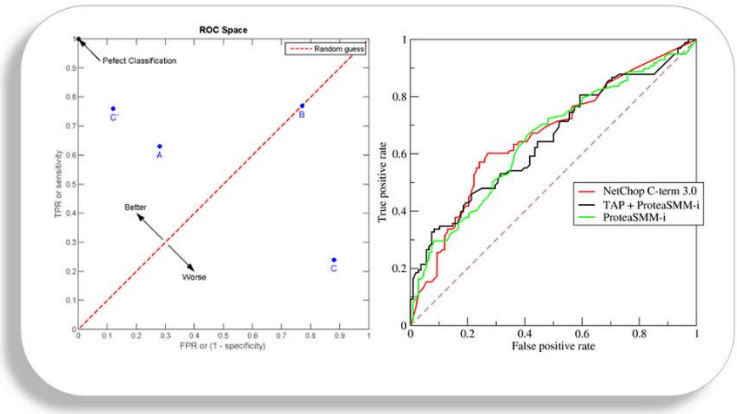


ROC curve

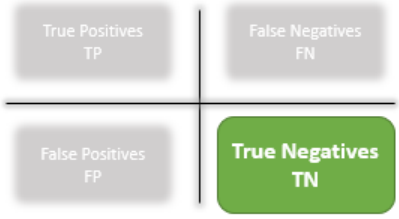
ROC stands for receiver operating characteristic and the graph is plotted against TPR and FPR for various threshold values. As TPR increases FPR also increases. As you can see in the first figure, we have four categories and we want the threshold value that leads us closer to the top left corner. Comparing different predictors (here 3) on a given dataset also becomes easy as you can see in figure 2, one can choose the threshold according to the application at hand. ROC AUC is just the area under the curve, the higher its numerical value the better.

$$True\ Positive\ Rate\ (TPR) = RECALL = \frac{TP}{TP+FN}$$

$$False\ Positive\ Rate\ (FPR) = 1 - Specificity = \frac{FP}{TN+FP}$$



PR vs ROC curve
Both the metrics are widely used to judge a models performance.
Which one to use PR or ROC?



The answer lies in TRUE NEGATIVES.
Due to the absence of TN in the precision-recall equation, they are useful in imbalanced classes. In the case of class imbalance when there is a majority of the negative class. The metric doesn't take much into consideration the high number of TRUE NEGATIVES of the negative class which is in majority, giving better resistance to the imbalance. This is important when the detection of the positive class is very important.
Like to detect cancer patients, which has a high class imbalance because very few have it out of all the diagnosed. We certainly don't want to miss on a person having cancer and going undetected (recall) and be sure the detected one is having it (precision).
Due to the consideration of TN or the negative class in the ROC equation, it is useful when both the classes are important to us. Like the detection of cats and dog. The importance of true negatives makes sure that both the classes are given importance, like the output of a ML/DL model in determining the image is of a cat or a dog.

Classification
Here the goal is to learn a mapping from inputs x to outputs y, where $y \in \{1,...,C\}$, with C being the number of classes. If C = 2, this is called binary classification (in which case we often assume $y \in \{0, 1\}$); if C > 2, this is called multiclass classification. If the class labels are not mutually exclusive (e.g., somebody may be classified as tall and strong), we call it multi-label classification, but this is best viewed as predicting multiple related binary class labels (a so-called multiple output model). When we use the term "classification", we will mean multiclass classification with a single output, unless we state otherwise.
One way to formalize the problem is as function approximation. We assume y = f(x) for some unknown function f, and the goal of learning is to estimate the function f given a labeled training set, and then to make predictions using $\hat{y} = \hat{f}(\mathbf{x})$. (We use the hat symbol to denote an estimate.) Our main goal is to make predictions on novel inputs, meaning ones that we have not seen before (this is called generalization), since predicting the response on the training set is easy.

LINEAR REGRESSION

Regression analysis may broadly be defined as the analysis of relationships among variables. This relationship is given as an equation that helps to predict the dependent variable Y through one or more independent variables. In regression analysis, the variable whose values vary with the variations in the values of the other variable(s) is called the **dependent variable** or **response variable**. The other variables which are independent in nature and influence the response variable are called **independent variables**, **predictor variables** or **regressor variables**.

Example: Suppose a statistician employed by a cold drink bottler is analysing the product delivery and service operation for vending machines. He would like to find how the delivery time taken by the delivery man to load and service a machine is related to the volume of delivery cases. The statistician visits 50 randomly chosen retailer shops having vending machines and observes the delivery time (in minutes) and the volume of delivery cases for each shop. He plots those 50 observations on a graph, which shows that an approximate linear relationship exists between the delivery time and delivery volume. If Y represents the delivery time and X, the delivery volume, the equation of a straight line relating these two variables may be given as

$$\mathbf{Y = a + bX \ldots (1)}$$

where a is the intercept and b, the slope.

In such cases, we draw a straight line in the form of equation (1) so that the data points generally fall near the straight line. Now, suppose the points do not fall exactly on the straight line. Then we should modify equation (1) to minimise the difference between the observed value of Y and that given by the straight line (a + b X). This is known as **error**.

The error e, which is the difference between the observed value and the predicted value of the variable of interest Y, may be conveniently assumed as a statistical error. This error term accounts for the variability in Y that cannot be explained by the linear relationship between X and Y. It may arise due to the effects of other factors. Thus, a more plausible model for the variable of interest (Y) may be given as

$$\mathbf{Y = a + b\,X + e \ldots (2)}$$

where the intercept a and the slope b are unknown constants and e is a random error component.

Equation (2) is called a **linear regression model**.

Fitting of regression line

Let the given data of n pairs of observations on X and Y be as follows:

X: $X_1\ X_2\ X_3$. … … Xi ….. … Xn

Y: $Y_1\ Y_2\ Y_3$. … … Yi …..… Yn

where Y is the dependent variable and X, the independent variable.

Suppose, we wish to fit the following simple regression equation to the data: Y = a + bX

where a is the intercept and b is the slope of the equation.

For fitting equation to the data on (X, Y), we follow the steps given below:

Step 1: We draw a scatter diagram by plotting the (X, Y) points given in data.

Step 2: We construct a table as given below and take the sum of the values of X_i, Y_i, X_iY_i, and X_i^2. We write the values of $\sum X$, $\sum Y$, $\sum XY$ and $\sum X^2$ in the last row.

Step 3: We express of aˆ given in equation (1) as follows:

$$\hat{a} = \bar{Y} - b\bar{X} = \frac{1}{n}\left[\sum Y - b\sum X\right]$$

Where

$$\hat{b} = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - \left(\sum X\right)^2}$$

substitute above values in the regression equation and get

$$\hat{Y} = \hat{a} + \hat{b}X$$

POLYNOMIAL REGRESSION

Some engineering data is poorly represented by a straight line. For these cases, a curve would be better suited to it these data. The method is to it polynomials to the data using polynomial regression.

Polynomial Regression is a form of regression analysis in which the relationship between the independent variables and dependent variables are modeled in the nth degree polynomial. Polynomial Regression models are usually fit with the method of least squares. The **least square method** minimizes the variance of the coefficients, under the Gauss Markov Theorem. Polynomial Regression is a special case of Linear Regression where we fit the polynomial equation on the data with a curvilinear relationship between the dependent and independent variables. A Quadratic Equation is a Polynomial Equation of 2nd Degree. However, this degree can increase to nth values.

The least-squares procedure can be readily extended to it the data to a higher-order polynomial. For example, suppose that we it a second-order polynomial or quadratic:

$$y = a_0 + a_1 x + a_2 x^2 + e$$

Where x-independent variable, y-dependent variable, a0,a1 and a2 are coefficients, and e –error term.

For this case the sum of the squares of the residuals is

$$S_r = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i - a_2 x_i^2)^2$$

we take the derivative with respect to each of the unknown coefficients of the polynomial, as in

$$\frac{\partial S_r}{\partial a_0} = -2 \sum (y_i - a_0 - a_1 x_i - a_2 x_i^2)$$

$$\frac{\partial S_r}{\partial a_1} = -2 \sum x_i (y_i - a_0 - a_1 x_i - a_2 x_i^2)$$

$$\frac{\partial S_r}{\partial a_2} = -2 \sum x_i^2 (y_i - a_0 - a_1 x_i - a_2 x_i^2)$$

These equations can be set equal to zero and rearranged to develop the following set of normal equations:

$$(n) a_0 + \left(\sum x_i\right) a_1 + \left(\sum x_i^2\right) a_2 = \sum y_i$$

$$\left(\sum x_i\right) a_0 + \left(\sum x_i^2\right) a_1 + \left(\sum x_i^3\right) a_2 = \sum x_i y_i$$

$$\left(\sum x_i^2\right) a_0 + \left(\sum x_i^3\right) a_1 + \left(\sum x_i^4\right) a_2 = \sum x_i^2 y_i$$

where all summations are from i = 1 through n. Note that the above three equations are linear and have three unknowns: a0 , a1 , and a2 . The coefficients of the unknowns can be calculated directly from the observed data. For this case, we see that the problem of determining a least-squares second-order polynomial is equivalent to solving a system of three simultaneous linear equations.

These equations can be rewritten in matrix form as follows:

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i^3 \\ \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i^3 & \sum_{i=1}^{n} x_i^4 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} x_i^2 y_i \end{bmatrix}$$

Where:

- *n* is the number of data points.
- *xi* and *yi* are the values of the independent and dependent variables for the *i*-th data point.

The most common way to solve a system of linear equations is by using matrix algebra and methods like Gaussian elimination or matrix inversion. Here are the general steps to solve a system of linear equations:

Step 1: Write Down the System of Equations

Write the system of linear equations in standard form, where each equation has the form:

$$a_1 x_1 + a_2 x_2 + \ldots + a_n x_n = b$$

Where:

- $a1,a2,\ldots,an$ are the coefficients of the variables $x1,x2,\ldots,xn$.
- b is the constant on the right-hand side of the equation.

You should have as many equations as there are variables.

Step 2: Create a Coefficient Matrix

Construct a coefficient matrix A by extracting the coefficients of the variables from the left side of each equation. This matrix will have dimensions $n \times n$ if you have n variables.

Step 3: Create a Right-Hand Side Vector

Create a right-hand side vector B by extracting the constants from the right side of each equation. This vector will have dimensions $n \times 1$.

Step 4: Solve the System

There are several methods to solve the system:

There are several methods to solve the system:

- **Matrix Inversion**: If A is invertible (non-singular), you can solve the system using the formula $X=A^{-1}B$, where X is the vector of solutions.
- **Gaussian Elimination**: Use row operations to transform the augmented matrix $[A|B]$ into row-echelon or reduced row-echelon form. Then, back-substitute to find the values of the variables.
- **Matrix Factorization**: Factorize A into LU or QR form and use the factorization to solve the system more efficiently.
- **Numerical Methods**: For large or ill-conditioned systems, numerical methods like the Gauss-Seidel method or the Conjugate Gradient method are used.

Here's a simple example using matrix inversion in Python:

```
import numpy as np
# Define the coefficient matrix A and the right-hand side vector B
A = np.array([[2, 1], [1, 3]])
B = np.array([4, 7])
# Solve for the variables X using matrix inversion
X = np.linalg.inv(A).dot(B)
print("Solution:")
print(X)
```

The above code solves the system $2x_1+x_2=4$ and $x_1+3x_2=7$ and prints the values of x_1 and x_2 as the solution.

Solution:

[1. 2.]

Polynomial regression is a type of regression analysis in which the relationship between the independent variable (X) and the dependent variable (Y) is modeled as an nth-degree polynomial. In Python, you can perform polynomial regression using libraries like NumPy and scikit-learn. Here's a basic example of polynomial regression using scikit-learn:

```
import numpy as np
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import PolynomialFeatures
import matplotlib.pyplot as plt

# Generate sample data
np.random.seed(0)
X = 2 * np.random.rand(100, 1)
Y = 4 + 3 * X + np.random.randn(100, 1)

# Fit a polynomial regression model
degree = 2  # You can change the degree as needed
poly_features = PolynomialFeatures(degree=degree)
X_poly = poly_features.fit_transform(X)
```

```
model = LinearRegression()
model.fit(X_poly, Y)

# Make predictions
X_new = np.linspace(0, 2, 100).reshape(-1, 1)
X_new_poly = poly_features.transform(X_new)
Y_new = model.predict(X_new_poly)

# Plot the original data and the polynomial regression curve
plt.scatter(X, Y, label='Original Data')
plt.plot(X_new, Y_new, 'r-', label='Polynomial Regression')
plt.xlabel('X')
plt.ylabel('Y')
plt.legend()
plt.show()
# The coefficients of the multivariate polynomial regression model
coefficients = model.coef_
intercept = model.intercept_
print("Coefficients:")
print(coefficients)
print("Intercept:")
print(intercept)
```

In this example:

1. We generate some sample data points with random noise.
2. We use **PolynomialFeatures** from scikit-learn to transform our input features **X** into polynomial features up to a specified degree.
3. We then fit a linear regression model to the polynomial features.
4. Finally, we make predictions using the model and plot the original data along with the polynomial regression curve.

You can adjust the **degree** variable to control the degree of the polynomial you want to fit to your data. Higher-degree polynomials can capture more complex relationships but may also lead to overfitting, so be cautious when choosing the degree.

MULTIPLE LINEAR REGRESSION

Multiple linear regression is a method we can use to quantify the relationship between two or more predictor variables and a response variable.

The Regression Line: With one independent variable, we may write the regression equation as:

$$Y = a + bX + e$$

Where Y is an observed score on the dependent variable, *a* is the intercept, *b* is the slope, X is the observed score on the independent variable, and *e* is an error or residual.

We can extend this to any number of independent variables:

$$Y = a + b_1X_1 + b_2X_2 + \ldots + b_kX_k + e \quad (3.1)$$

Note that we have *k* independent variables and a slope for each. We still have one error and one intercept. Again we want to choose the estimates of *a* and *b* so as to minimize the sum of squared errors of prediction. The prediction equation is:

$$Y' = a + b_1X_1 + b_2X_2 + \ldots + b_kX_k \quad (3.2)$$

Finding the values of *b* (the slopes) is tricky for k>2 independent variables, and you really need matrix algebra to see the computations. It's simpler for k=2 IVs, which we will discuss here. But the basic ideas are the same no matter how many independent variables you have. If you understand the meaning of the slopes with two independent variables, you will likely be good no matter how many you have.

For the one variable case, the calculation of *b* and *a* was:

$$b = \frac{\sum xy}{\sum x^2}$$

$$a = \overline{Y} - b\overline{X}$$

For the two variable case:

$$b_1 = \frac{(\sum x_2^2)(\sum x_1 y) - (\sum x_1 x_2)(\sum x_2 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

and

$$b_2 = \frac{(\sum x_1^2)(\sum x_2 y) - (\sum x_1 x_2)(\sum x_1 y)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

where

- $\Sigma x_1^2 = \Sigma X_1^2 - ((\Sigma X_1)^2 / n)$
- $\Sigma x_2^2 = \Sigma X_2^2 - ((\Sigma X_2)^2 / n)$
- $\Sigma x_1 y = \Sigma X_1 y - ((\Sigma X_1 \Sigma y) / n)$
- $\Sigma x_2 y = \Sigma X_2 y - ((\Sigma X_2 \Sigma y) / n)$
- $\Sigma x_1 x_2 = \Sigma X_1 X_2 - ((\Sigma X_1 \Sigma X_2) / n)$

At this point, you should notice that all the terms from the one variable case appear in the two variable case. In the two variable case, the other X variable also appears in the equation. For example, X_2 appears in the equation for b_1. Note that terms corresponding to the variance of both X variables occur in the slopes. Also note that a term corresponding to the covariance of X1 and X2 (sum of deviation cross-products) also appears in the formula for the slope.
The equation for *a* with two independent variables is:

$$a = \overline{Y} - b_1\overline{X}_1 - b_2\overline{X}_2$$

This equation is a straight-forward generalization of the case for one independent variable.

MLR MATRIX FORMULATION

Multiple linear regression is a generalized form of simple linear regression, in which the data contains multiple explanatory variables.

	SLR		MLR				
	x	**y**	$\mathbf{x}_1$	$\mathbf{x}_2$	...	$\mathbf{x}_p$	**y**
case 1:	x_1	y_1	x_{11}	x_{12}	...	x_{1p}	y_1
case 2:	x_2	y_2	x_{21}	x_{22}	...	x_{2p}	y_2
	⋮	⋮	⋮	⋮	⋱	⋮	⋮
case *n*:	x_n	y_n	x_{n1}	x_{n2}	...	x_{np}	y_n

- For SLR, we observe pairs of variables.
- For MLR, we observe rows of variables.
- Each row (or pair) is called a case, a record, or a data point
- yi is the response (or dependent variable) of the ith observation
- There are p explanatory variables (or covariates, predictors, independent variables), and xik is the value of the explanatory variable xk of the ith case

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \varepsilon_i \quad \text{where } \varepsilon_i\text{'s are i.i.d. } N(0, \sigma^2)$$

In the model above,

- ε_i's (errors, or noise) are i.i.d. $N(0, \sigma^2)$
- Parameters include:

$\beta_0 =$ intercept;

$\beta_k =$ regression coefficients (slope) for the kth explanatory variable, $k = 1, \ldots, p$

$\sigma^2 = \mathrm{Var}(\varepsilon_i)$ is the variance of errors

Refer further https://online.stat.psu.edu/stat462/node/132/

REGRESSION METRICS

Residual

The difference between the fitted value $\hat{Y}_i$ and Yi is known as the residual and is denoted by

$r_i = Y_i - \hat{Y}_i, \qquad i = 1, 2, \ldots, n$

The role of the residuals and its analysis is very important in regression modelling.

Mean Squared Error (MSE)

Mean squared error (MSE) measures the amount of error in statistical models. It assesses the average squared difference between the observed and predicted values. When a model has no error, the MSE equals zero. As model error increases, its value increases. The mean squared error is also known as the ***mean squared deviation (MSD)***.

For example, in regression, the mean squared error represents the average squared residual/error.

$$MSE = \frac{\sum(y_i - \hat{y}_i)^2}{n}$$

Where:
yi is the ith observed value.
ŷi is the corresponding predicted value.
n = the number of observations.

Squaring the error gives higher weight to the outliers, which results in a smooth gradient for small errors. Optimization algorithms benefit from this penalization for large errors as it is helpful in finding the optimum values for parameters. MSE will never be negative since the errors are squared. The value of the error ranges from zero to infinity. MSE increases exponentially with an increase in error. A good model will have an MSE value closer to zero.

Root Mean Square Error (RMSE)

Root Mean Squared Error (RMSE) is a popular metric used in machine learning and statistics to measure the accuracy of a predictive model. It quantifies the differences between predicted values and actual values, squaring the errors, taking the mean, and then finding the square root. RMSE provides a clear understanding of the model's performance, with lower values indicating better predictive accuracy.

RMSE is computed by taking the square root of MSE. RMSE is also called the Root Mean Square Deviation. It measures the average magnitude of the errors and is concerned with the deviations from the actual value. RMSE value with zero indicates that the model has a perfect fit. The lower the RMSE, the better the model and its predictions. A higher RMSE indicates that there is a large deviation from the residual to the ground truth. RMSE can be used with different features as it helps in figuring out if the feature is improving the model's prediction or not.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Ref: https://scikit-learn.org/stable/modules/model_evaluation.html#regression-metrics

Questions:

1. The data set of size n = 15 (Yield data) contains measurements of yield from an experiment done at five different temperature levels. The variables are y = yield and x = temperature in degrees Fahrenheit. The table below gives the data used for this analysis.

i	Temp.	Yield
1	50	3.3
2	50	2.8
3	50	2.9
4	70	2.3
5	70	2.6
6	70	2.1
7	80	2.5
8	80	2.9
9	80	2.4
10	90	3.0
11	90	3.1
12	90	2.8
13	100	3.3
14	100	3.5
15	100	3.0

a. Create a CSV file with sample data.
b. Write a Python function program to:
Find the fitted simple linear and polynomial regression equations for the given data.
c. Compare the coefficients obtained from manually intuitive and matrix formulation methods with your program.
d. Plot the scatterplot of the raw data and then another scatterplot with lines pertaining to a linear fit and a quadratic fit overlayed.
e. Compute the error, MSE, and RMSE.
Note: Do not use scikit-learn.

2. When heart muscle is deprived of oxygen, the tissue dies and leads to a heart attack ("myocardial infarction"). Apparently, cooling the heart reduces the size of the heart attack. It is not known, however, whether cooling is only effective if it takes place before the blood, flow to the heart becomes restricted. Some researchers (Hale, et al, 1997) hypothesized that cooling the heart would be effective in reducing the size of the heart attack even if it takes place after the blood flow becomes restricted.
To investigate their hypothesis, the researchers conducted an experiment on 32 anesthetized rabbits that were subjected to a heart attack. The researchers established three experimental groups:

- *Rabbits whose hearts were cooled to 6º C within 5 minutes of the blocked artery ("**early cooling**")*
- *Rabbits whose hearts were cooled to 6º C within 25 minutes of the blocked artery ("**late cooling**")*
- *Rabbits whose hearts were not cooled at all ("**no cooling**")*

*At the end of the experiment, the researchers measured the size of the **infarcted** (i.e., damaged) **area** (in grams) in each of the 32 rabbits. But, as you can imagine, there is great variability in the size of hearts. The size of a rabbit's infarcted area may be large only because it has a larger heart. Therefore, in order to adjust for differences in heart sizes, the researchers also measured the size of the **region at risk** for infarction (in grams) in each of the 32 rabbits.*

Infarc	***Area***	***Group***	***X2***	***X3***
0.119	*0.34*	*3*	*0*	*0*
0.19	*0.64*	*3*	*0*	*0*
0.395	*0.76*	*3*	*0*	*0*
0.469	*0.83*	*3*	*0*	*0*
0.13	*0.73*	*3*	*0*	*0*
0.311	*0.82*	*3*	*0*	*0*
0.418	*0.95*	*3*	*0*	*0*

0.48	*1.06*	*3*	*0*	*0*
0.687	*1.2*	*3*	*0*	*0*
0.847	*1.47*	*3*	*0*	*0*
0.062	*0.44*	*1*	*1*	*0*
0.122	*0.77*	*1*	*1*	*0*
0.033	*0.9*	*1*	*1*	*0*
0.102	*1.07*	*1*	*1*	*0*
0.206	*1.01*	*1*	*1*	*0*
0.249	*1.03*	*1*	*1*	*0*
0.22	*1.16*	*1*	*1*	*0*
0.299	*1.21*	*1*	*1*	*0*
0.35	*1.2*	*1*	*1*	*0*
0.35	*1.22*	*1*	*1*	*0*
0.588	*0.99*	*1*	*1*	*0*
0.379	*0.77*	*2*	*0*	*1*
0.149	*1.05*	*2*	*0*	*1*
0.316	*1.06*	*2*	*0*	*1*
0.39	*1.02*	*2*	*0*	*1*
0.429	*0.99*	*2*	*0*	*1*
0.477	*0.97*	*2*	*0*	*1*
0.439	*1.12*	*2*	*0*	*1*
0.446	*1.23*	*2*	*0*	*1*
0.538	*1.19*	*2*	*0*	*1*
0.625	*1.22*	*2*	*0*	*1*
0.974	*1.4*	*2*	*0*	*1*

With their measurements in hand, the researchers' primary research question was: Does the mean size of the infarcted area differ among the three treatment groups — no cooling, early cooling, and late cooling — when controlling for the size of the region at risk for infarction?
A regression model that the researchers might use in answering their research question is:

$$y_i=(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\beta_3 x_{i3})+\epsilon_i$$

where:

- y_i *is the size of the infarcted area (in grams) of rabbit i*
- x_{i1} *is the size of the region at risk (in grams) of rabbit i*
- x_{i2} *= 1 if early cooling of rabbit i, 0 if not*
- x_{i3} *= 1 if late cooling of rabbit i, 0 if not*

and the independent error terms ε_i *follow a normal distribution with mean 0 and equal variance* σ^2*. Illustrates the need for being able to "translate" a research question into a statistical procedure. Often, the procedure involves four steps, namely:*

- *formulating a multiple regression model*
- *determining how the model helps answer the research question*
- *checking the model*
- *and performing a hypothesis test (or calculating a confidence interval)*

a. Create a CSV file with sample data.
b. Write a Python function program to:
c. Find the fitted multiple linear regression equation for the given data.
d. Compare the coefficients obtained manually using intuitive and matrix formulation methods with your program.
e. Plot the data adorned with the estimated regression equation.
f. Compute the error, MSE, and RMSE.
Note: Do not use scikit-learn.

Additional questions

1. Consider the Table Contains the Average Annual Gold Rate from 1965 – 2022. Gold prices fluctuated throughout the year 2020 because of the COVID-19 epidemic. With gold functioning as a safe haven for investors, demand for the precious metal grew, and its price followed suit. During the epidemic, the stock market weakened, but it began to recover by the end of 2020 when the price of gold fell slightly.

It's crucial to remember that gold prices fluctuate during the year, and the figure below represents the average price for that year.

With the exception of a few lows shared across a few years, The table shows that the gold price trend has always been upward, supporting the claim that gold is a secure investment over extended periods of time.

Create CSV file and Write a python program to find the fitted simple linear regression equation for the given data. Compare the coefficients obtained from sklearn model with your program. Compute the error, MSE and RMSE. Predict the gold price with the year 2025 for 1 gram.

This Table Contains the Average Annual Gold Rate from 1965 - 2022			
Year	**Price (24 karat per 10 grams)**	**Year**	**Price (24 karat per 10 grams)**
2022	₹ 52,950	1993	₹ 4,140
2021	₹ 50,045	1992	₹ 4,334
2020	₹ 48,651	1991	₹ 3,466
2019	₹ 35,220	1990	₹ 3,200
2018	₹ 31,438	1989	₹ 3,140
2017	₹ 29,667	1988	₹ 3,130
2016	₹ 28,623	1987	₹ 2,570
2015	₹ 26,343	1986	₹ 2,140
2014	₹ 28,006	1985	₹ 2,130
2013	₹ 29,600	1984	₹ 1,970
2012	₹ 31,050	1983	₹ 1,800
2011	₹ 26,400	1982	₹ 1,645
2010	₹ 18,500	1981	₹ 1,800
2009	₹ 14,500	1980	₹ 1,330
2008	₹ 12,500	1979	₹ 937
2007	₹ 10,800	1978	₹ 685
2006	₹ 8,400	1977	₹ 486

2005	₹ 7,000	1976	₹ 432
2004	₹ 5,850	1975	₹ 540
2003	₹ 5,600	1974	₹ 506
2002	₹ 4,990	1973	₹ 279
2001	₹ 4,300	1972	₹ 202
2000	₹ 4,400	1971	₹ 193
1999	₹ 4,234	1970	₹ 184
1998	₹ 4,045	1969	₹ 176
1997	₹ 4,725	1968	₹ 162
1996	₹ 5,160	1967	₹ 103
1995	₹ 4,680	1966	₹ 84
1994	₹ 4,598	1965	₹ 72

2. Consider the Question no 1 gold price with following year-wise silver price. Create a CSV file and Write a python program to find the fitted multiple linear regression equation for the given data. Compare the coefficients obtained from sklearn model with your program. Compute the error, MSE and RMSE. Predict the gold and silver price with the year 2024 for 1 gram.

Year	**Silver Rates in Rs./Kg.**	**Year**	**Silver Rates in Rs./Kg.**
1981	Rs.2715	2002	Rs.7875
1982	Rs.2720	2003	Rs.7695
1983	Rs.3105	2004	Rs.11770
1984	Rs.3570	2005	Rs.10675
1985	Rs.3955	2006	Rs.17405
1986	Rs.4015	2007	Rs.19520
1987	Rs.4794	2008	Rs.23625
1988	Rs.6066	2009	Rs.22165
1989	Rs.6755	2010	Rs.27255
1990	Rs.6463	2011	Rs.56900
1991	Rs.6646	2012	Rs.56290
1992	Rs.8040	2013	Rs.54030
1993	Rs.5489	2014	Rs.43070
1994	Rs.7124	2015	Rs.37825
1995	Rs.6335	2016	Rs.36990
1996	Rs.7346	2017	Rs.37825
1997	Rs.7345	2018	Rs.41400
1998	Rs.8560	2019	Rs.40600
1999	Rs.7615	2020	Rs.63435
2000	Rs.7900	2021	Rs.62572
2001	Rs.7215	2022	Rs.55100

3. Suppose you have a gold/silver price dataset with a single independent variable (X) and a dependent variable (Y). You want to fit a polynomial regression model to this data. Implement the process of selecting the appropriate degree for the polynomial (e.g., linear, quadratic, cubic) based on the dataset using Python.

4. Imagine you have a gold and silver price dataset with two independent variables (X1 and X2) and a dependent variable (Y). Implement in python, how you can perform multivariate polynomial regression to model the relationship between the independent variables and the dependent variable.

WEEK -05: LOGISTIC REGRESSION AND STOCHASTIC GRADIENT DESCENT (SGD)

Logistic regression is a supervised machine learning algorithm mainly used for classification tasks where the goal is to predict the probability that an instance of belonging to a given class or not. It is a kind of statistical algorithm, which analyze the relationship between a set of independent variables and the dependent binary variables. It is a powerful tool for decision-making. For example email spam or not.

It is used for classification algorithms its name is logistic regression. it's referred to as regression because it takes the output of the linear regression function as input and uses a sigmoid function to estimate the probability for the given class. The difference between linear regression and logistic regression is that linear regression output is the continuous value that can be anything while logistic regression predicts the probability that an instance belongs to a given class or not.

It is used for predicting the categorical dependent variable using a given set of independent variables.

- Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value.
- It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.
- Logistic Regression is much similar to the Linear Regression except that how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving the classification problems.
- In Logistic regression, instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).
- The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc.
- Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets.
- Logistic Regression can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification.

Logistic Function (Sigmoid Function):

- The sigmoid function is a mathematical function used to map the predicted values to probabilities.
- It maps any real value into another value within a range of 0 and 1. o The value of the logistic regression must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the "S" form.
- The S-form curve is called the Sigmoid function or the logistic function.
- In logistic regression, we use the concept of the threshold value, which defines the probability of either 0 or 1. Such as values above the threshold value tends to 1, and a value below the threshold values tends to 0.

Type of Logistic Regression:

On the basis of the categories, Logistic Regression can be classified into three types:

1. **Binomial:** In binomial Logistic regression, there can be only two possible types of the dependent variables, such as 0 or 1, Pass or Fail, etc.
2. **Multinomial:** In multinomial Logistic regression, there can be 3 or more possible unordered types of the dependent variable, such as "cat", "dogs", or "sheep"
3. **Ordinal:** In ordinal Logistic regression, there can be 3 or more possible ordered types of dependent variables, such as "low", "Medium", or "High".

Terminologies involved in Logistic Regression:

Here are some common terms involved in logistic regression:

- **Independent variables:** The input characteristics or predictor factors applied to the dependent variable's predictions.
- **Dependent variable:** The target variable in a logistic regression model, which we are trying to predict.
- **Logistic function:** The formula used to represent how the independent and dependent variables relate to one another. The logistic function transforms the input variables into a probability value between 0 and 1, which represents the likelihood of the dependent variable being 1 or 0.
- **Odds:** It is the ratio of something occurring to something not occurring. it is different from probability as the probability is the ratio of something occurring to everything that could possibly occur.
- **Log-odds:** The log-odds, also known as the logit function, is the natural logarithm of the odds. In logistic regression, the log odds of the dependent variable are modeled as a linear combination of the independent variables and the intercept.

- **Coefficient:** The logistic regression model's estimated parameters, show how the independent and dependent variables relate to one another.
- **Intercept:** A constant term in the logistic regression model, which represents the log odds when all independent variables are equal to zero.
- **Maximum likelihood estimation:** The method used to estimate the coefficients of the logistic regression model, which maximizes the likelihood of observing the data given the model.

How does Logistic Regression work?

The logistic regression model transforms the linear regression function continuous value output into categorical value output using a sigmoid function, which maps any real-valued set of independent variables input into a value between 0 and 1. This function is known as the logistic function.

Let the independent input features be

$$X = \begin{bmatrix} x_{11} & \dots & x_{1m} \\ x_{21} & \dots & x_{2m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \dots & x_{nm} \end{bmatrix}$$

and the dependent variable is Y having only binary value i.e. 0 or 1.

$$Y = \begin{cases} 0 & \text{if } Class\ 1 \\ 1 & \text{if } Class\ 2 \end{cases}$$

then apply the multi-linear function to the input variables X

$$z = (\sum_{i=1}^{n} w_i x_i) + b$$

x_i $w_i = [w_1, w_2, w_3, \cdots, w_m]$

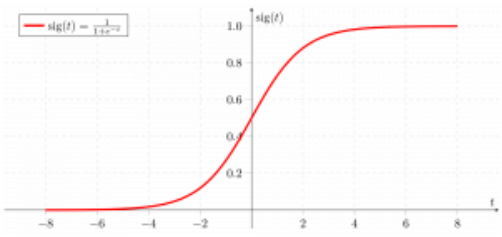
Here is the ith observation of X, is the weights or Coefficient, and b is the bias term also known as intercept. simply this can be represented as the dot product of weight and bias.

$$z = w \cdot X + b$$

Sigmoid Function

Now we use the sigmoid function where the input will be z and we find the probability between 0 and 1. i.e predicted y.

$$\sigma(z) = \frac{1}{1-e^{z}}$$



Sigmoid function

As shown above, the figure sigmoid function converts the continuous variable data into the probability i.e. between 0 and 1.

- $\sigma(z)$ tends towards 1 as $z \to \infty$
- $\sigma(z)$ tends towards 0 as $z \to -\infty$
- $\sigma(z)$ is always bounded between 0 and 1

where the probability of being a class can be measured as:

$$P(y=1) = \sigma(z)$$
$$P(y=0) = 1 - \sigma(z)$$

Logistic Regression Equation

The odd is the ratio of something occurring to something not occurring. it is different from probability as the probability is the ratio of something occurring to everything that could possibly occur. so odd will be

$$\frac{p(x)}{1-p(x)} = e^z$$

Applying natural log on odd. then log odd will be

$$\log\left[\frac{p(x)}{1-p(x)}\right] = z$$
$$\log\left[\frac{p(x)}{1-p(x)}\right] = w \cdot X + b$$

then the final logistic regression equation will be:

$$p(X; b, w) = \frac{e^{w \cdot X + b}}{1 + e^{w \cdot X + b}} = \frac{1}{1 + e^{-w \cdot X + b}}$$

Likelihood function for Logistic Regression

The predicted probabilities will p(X;b,w) = p(x) for y=1 and for y = 0 predicted probabilities will 1-p(X;b,w) = 1-p(x)

$$L(b, w) = \prod_{i=1} np(x_i)^{y_i}(1 - p(x_i))^{1-y_i}$$

Taking natural logs on both sides

$$\begin{aligned} l(b, w) = \log(L(b, w)) &= \sum_{i=1}^{n} y_i \log p(x_i) \; + \; (1 - y_i) \log(1 - p(x_i)) \\ &= \sum_{i=1}^{n} y_i \log p(x_i) + \log(1 - p(x_i)) - y_i \log(1 - p(x_i)) \\ &= \sum_{i=1}^{n} \log(1 - p(x_i)) + \sum_{i=1}^{n} y_i \log \frac{p(x_i)}{1 - p(x_i} \\ &= \sum_{i=1}^{n} - \log 1 - e^{-(w \cdot x_i + b)} + \sum_{i=1}^{n} y_i (w \cdot x_i + b) \\ &= \sum_{i=1}^{n} - \log 1 + e^{w \cdot x_i + b} + \sum_{i=1}^{n} y_i (w \cdot x_i + b) \end{aligned}$$

Gradient of the log-likelihood function

To find the maximum likelihood estimates, we differentiate w.r.t w,

$$\begin{aligned}\frac{\partial J(l(b,w)}{\partial w_j} &= -\sum_{i=n}^{n} \frac{1}{1+e^{w\cdot x_i+b}} e^{w\cdot x_i+b} x_{ij} + \sum_{i=1}^{n} y_i x_{ij} \\ &= -\sum_{i=n}^{n} p(x_i;b,w) x_{ij} + \sum_{i=1}^{n} y_i x_{ij} \\ &= \sum_{i=n}^{n} (y_i - p(x_i;b,w)) x_{ij}\end{aligned}$$

Assumptions for Logistic Regression

The assumptions for Logistic regression are as follows:

- **Independent observations:** Each observation is independent of the other. meaning there is no correlation between any input variables.
- **Binary dependent variables:** It takes the assumption that the dependent variable must be binary or dichotomous, meaning it can take only two values. For more than two categories softmax functions are used.
- **Linearity relationship between independent variables and log odds:** The relationship between the independent variables and the log odds of the dependent variable should be linear.
- **No outliers:** There should be no outliers in the dataset.
- **Large sample size:** The sample size is sufficiently large

Logistic regression is a commonly used algorithm for binary classification problems. Here's a Python program for logistic regression with an example using the popular Iris dataset for classification:

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn import datasets
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import classification_report, confusion_matrix

# Load the Iris dataset
iris = datasets.load_iris()
X = iris.data[:, :2]  # We'll use only the first two features for simplicity
y = (iris.target != 0) * 1  # Convert target labels to binary (0 or 1)

# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

# Standardize the feature data
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)

# Create a logistic regression model
model = LogisticRegression(solver='liblinear')

# Train the model
model.fit(X_train, y_train)

# Make predictions on the test set
y_pred = model.predict(X_test)

# Evaluate the model
print("Confusion Matrix:")
```

```
print(confusion_matrix(y_test, y_pred))

print("\nClassification Report:")
print(classification_report(y_test, y_pred))

# Plot the decision boundary
x_min, x_max = X[:, 0].min() - 1, X[:, 0].max() + 1
y_min, y_max = X[:, 1].min() - 1, X[:, 1].max() + 1
xx, yy = np.meshgrid(np.arange(x_min, x_max, 0.01), np.arange(y_min, y_max, 0.01))
Z = model.predict(np.c_[xx.ravel(), yy.ravel()])
Z = Z.reshape(xx.shape)

plt.contourf(xx, yy, Z, cmap=plt.cm.RdBu, alpha=0.8)
plt.scatter(X[:, 0], X[:, 1], c=y, cmap=plt.cm.RdBu)
plt.xlabel('Sepal Length (standardized)')
plt.ylabel('Sepal Width (standardized)')
plt.title('Logistic Regression Decision Boundary')
plt.show()
```

In this program:

1. We load the Iris dataset and select only the first two features for binary classification.
2. We split the data into training and testing sets using **train_test_split**.
3. We standardize the feature data using **StandardScaler** to have zero mean and unit variance.
4. We create a logistic regression model using **LogisticRegression** from scikit-learn.
5. We train the model on the training data.
6. We make predictions on the test data and evaluate the model's performance using a confusion matrix and a classification report.
7. Finally, we plot the decision boundary of the logistic regression model to visualize how it separates the two classes.

You can adjust the dataset, features, and model parameters as needed for your specific classification problem.

Stochastic gradient descent (SGD)

Ref: https://machinelearningmastery.com/linear-regression-tutorial-using-gradient-descent-for-machine-learning/

Questions

Note: "Refer to the table that contains the average annual gold rate from 1965 to 2022 and the year-wise silver prices available at Week-4 ML lab manual."

1. Let's say we have a fictional dataset of pairs of variables, a mother and her daughter's heights:

mother height	daughter height
58	60
62	60
60	58
64	60
67	70
70	72

height of mother(x)/daughter (y) pairs

Create a CSV file for the above training data and write a Python function program to find the fitted linear regression with gradient descent technique. Compare the coefficients obtained from the sklearn model with your program. Compute the error, MSE and RMSE. Plot the graph Daughter height (Y-axis) vs Mother height (X-axis) with blue colour. Also, plot the line of best fit with red colour. Predict her daughter's height with given a new mother height as 63. Plot the graph of error in y-axis and iteration in x-axis with 4 epochs (6x4=24 iterations).

2.

Hours of Study (X)	Pass (Y)
1	0
2	0
3	0
4	0
5	1
6	1
7	1
8	1

Here, X is the number of hours of study, and Y is the outcome (0 for fail, 1 for pass).

Create a CSV file for the above training data and write a Python function program to find the fitted logistic regression with gradient descent technique. Compare the coefficients obtained from the sklearn model with your program. Compute the predicted y and assign the class label (prediction = 0 IF p(fail) < 0.5 and prediction = 1 IF p(pass) >= 0.5) and compute the accuracy. Find the error for each iteration and predict the probability that a student will pass the exam if they study for a) 3.5 hours b) 7.5 hours. Plot the graph of error in y-axis and iteration in x-axis with 3 epochs (8x3=24 iterations).

3.

	x_1	x_2	y
1)	4	1	2
2)	2	8	-14
3)	1	0	1
4)	3	2	-1
5)	1	4	-7
6)	6	7	-8

Consider the above dataset with two independent variables (X1 and X2) and a dependent variable (Y). Implement in python, how you can perform the logistic regression to model the relationship between the independent variables and the dependent variable.

Additional Questions

1. Write a python program for SGD by considering the year wise gold and silver price data. Compare the coefficients obtained from sklearn model with your program. Compute the error, MSE and RMSE. Predict the gold price with the year 2025 for 1 gram and, gold and silver price with the year 2024 for 1 gram.

2. Suppose you have a gold/silver price dataset with a single independent variable (X) and a dependent variable (Y). You want to fit a logistic regression model to this data. Develop an example code snippet in Python.

3. Consider the gold and/or silver price dataset and to evaluate a logistic regression model using ROC and AUC:

1. Calculate predicted probabilities for each instance in the test set.
2. Plot the ROC curve using the True Positive Rate (Sensitivity) and False Positive Rate.
3. Calculate the AUC to summarize the model's performance.

WEEK-07: NAÏVE BAYES CLASSIFIER

BAYES' THEOREM

Bayes' theorem describes the probability of occurrence of an event related to any condition. It is also considered for the case of conditional probability. Bayes theorem is also known as the formula for the probability of "causes". For example: if we have to calculate the probability of taking a blue ball from the second bag out of three different bags of balls, where each bag contains three different colour balls viz. red, blue, black. In this case, the probability of occurrence of an event is calculated depending on other conditions is known as conditional probability.

Bayes Theorem Statement

Let E_1, E_2,…, E_n be a set of events associated with a sample space S, where all the events E_1, E_2,…, E_n have nonzero probability of occurrence and they form a partition of S. Let A be any event associated with S, then according to Bayes theorem,

$$P(E_i \mid A) = \frac{P(E_i)P(A \mid E_i)}{\sum_{k=1}^{n} P(E_k)P(A|E_k)}$$

for any k = 1, 2, 3, ..., n

Bayes Theorem Proof

According to the conditional probability formula,

$$P(E_i \mid A) = \frac{P(E_i \cap A)}{P(A)} \ldots (1)$$

Using the multiplication rule of probability,

$$P(E_i \cap A) = P(E_i)P(A \mid E_i) \ldots (2)$$

Using total probability theorem,

$$P(A) = \sum_{k=1}^{n} P(E_k)P(A|E_k) \ldots (3)$$

Putting the values from equations (2) and (3) in equation 1, we get

$$P(E_i \mid A) = \frac{P(E_i)P(A \mid E_i)}{\sum_{k=1}^{n} P(E_k)P(A|E_k)}$$

Note:

The following terminologies are also used when the Bayes theorem is applied:

Hypotheses: The events E_1, E_2,… E_n is called the hypotheses

Priori Probability: The probability $P(E_i)$ is considered as the priori probability of hypothesis E_i

Posteriori Probability: The probability $P(E_i|A)$ is considered as the posteriori probability of hypothesis E_i

Bayes' theorem is also called the formula for the probability of "causes". Since the E_i's are a partition of the sample space S, one and only one of the events E_i occurs (i.e. one of the events E_i must occur and the only one can occur). Hence, the above formula gives us the probability of a particular E_i (i.e. a "Cause"), given that the event A has occurred.

Bayes Theorem Derivation

Bayes Theorem can be derived for events and random variables separately using the definition of conditional probability and density.

From the definition of conditional probability, Bayes theorem can be derived for events as given below:

$P(A|B) = P(A \cap B)/ P(B)$, where $P(B) \neq 0$

$P(B|A) = P(B \cap A)/ P(A)$, where $P(A) \neq 0$

Here, the joint probability $P(A \cap B)$ of both events A and B being true such that,

$P(B \cap A) = P(A \cap B)$

$P(A \cap B) = P(A \mid B)\, P(B) = P(B \mid A)\, P(A)$

$P(A|B) = [P(B|A)\, P(A)]/ P(B)$, where $P(B) \neq 0$

Naming the Terms in the Theorem
The terms in the Bayes Theorem equation are given names depending on the context where the equation is used. It can be helpful to think about the calculation from these different perspectives and help to map your problem onto the equation.

Firstly, in general, the result P(A|B) is referred to as the posterior probability and P(A) is referred to as the prior probability.
P(A|B): Posterior probability.
P(A): Prior probability.
Sometimes P(B|A) is referred to as the likelihood and P(B) is referred to as the evidence.
P(B|A): Likelihood.
P(B): Evidence.
This allows Bayes Theorem to be restated as:
Posterior = Likelihood * Prior / Evidence

Implementation in python
```
# Define the prior probability P(A), likelihood P(B|A), and evidence P(B)
prior_probability = 0.01  # P(A) - Prior probability of event A
likelihood = 0.9  # P(B|A) - Likelihood of observing event B given A
evidence = 0.02  # P(B) - Total probability of observing event B

# Calculate the posterior probability P(A|B) using Bayes' theorem
posterior_probability = (prior_probability * likelihood) / evidence

# Print the result
print(f"Posterior Probability P(A|B): {posterior_probability:.4f}")
```

Example
Suppose the probability of the weather being cloudy is 40%. Also suppose the probability of rain on a given day is 20%. Also suppose the probability of clouds on a rainy day is 85%. If it's cloudy outside on a given day, what is the probability that it will rain that day?
Solution:
- P(cloudy) = 0.40
- P(rain) = 0.20
- P(cloudy | rain) = 0.85

Thus, we can calculate:
- P(rain | cloudy) = P(rain) * P(cloudy | rain) / P(cloudy)
- P(rain | cloudy) = 0.20 * 0.85 / 0.40
- P(rain | cloudy) = 0.425

Implementation in python
```
#define function for Bayes' theorem
def bayesTheorem(pA, pB, pBA):
    return pA * pBA / pB
#define probabilities
pRain = 0.2
pCloudy = 0.4
pCloudyRain = 0.85
#use function to calculate conditional probability
bayesTheorem(pRain, pCloudy, pCloudyRain)
```

NAÏVE BAYES CLASSIFIER

The Naive Bayes classifier is a simple yet powerful machine learning algorithm that is particularly well-suited for tasks involving text classification and spam filtering. It is based on Bayes' theorem and makes the "naive" assumption of independence between the features used for classification. Despite this simplification, Naive Bayes often performs surprisingly well in practice. In this analysis, we will delve into the key concepts behind the Naive Bayes classifier, its applications, advantages, limitations, and practical implementation.

Key Concepts of Naive Bayes Classifier

1. **Bayes' Theorem**: At the heart of the Naive Bayes classifier is Bayes' theorem, a fundamental concept in probability theory. It allows us to calculate the probability of a particular event based on prior knowledge of conditions that might be related to the event.

Bayes' theorem can be expressed as:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

Where:

- P(A|B) is the conditional probability of event A given event B.
- P(B|A) is the conditional probability of event B given event A.
- P(A) and P(B) are the probabilities of events A and B, respectively.

2. **Independence Assumption**: The "naive" part of Naive Bayes comes from assuming that the features used for classification are independent of each other. In practice, this assumption is often not entirely accurate, but the model can still perform well.
3. **Multinomial and Gaussian Naive Bayes**: There are different variants of Naive Bayes classifiers, including Multinomial and Gaussian Naive Bayes. Multinomial Naive Bayes is commonly used for text data, while Gaussian Naive Bayes is suitable for continuous data with a normal distribution.

Applications of Naive Bayes Classifier

The Naive Bayes classifier finds applications in various domains:

1. **Text Classification**: It is widely used for spam email detection, sentiment analysis, and topic classification of documents.
2. **Document Categorization**: Naive Bayes can categorize documents into predefined categories, making it useful for news article classification and content recommendation.
3. **Medical Diagnosis**: It can assist in medical diagnosis by classifying patient data into different disease categories.
4. **Customer Sentiment Analysis**: Businesses use it to analyze customer feedback, reviews, and social media comments to determine customer sentiment.
5. **Recommendation Systems**: Naive Bayes can be used to recommend products or content based on user behavior and preferences.

Advantages of Naive Bayes Classifier

1. **Simplicity**: Naive Bayes is easy to understand and implement, making it a good choice for quick prototyping and initial model development.
2. **Efficiency**: It is computationally efficient and can handle large datasets with a relatively small amount of memory.
3. **Good for Text Data**: Naive Bayes performs well on text data, making it a popular choice for tasks like spam detection and sentiment analysis.
4. **Works with Small Data**: Even with limited training data, Naive Bayes can produce meaningful results.

Limitations of Naive Bayes Classifier

1. **Independence Assumption**: The assumption of feature independence is often violated in real-world data, which can lead to suboptimal performance.
2. **Limited Expressiveness**: Naive Bayes is a linear classifier, which means it cannot capture complex relationships between features.
3. **Poor with Rare Events**: It may perform poorly when dealing with rare events or classes with limited examples.
4. **Sensitive to Feature Quality**: The quality of features used in the model can greatly impact its performance.

Practical Implementation of Naive Bayes Classifier

Implementing a Naive Bayes classifier typically involves the following steps:

1. **Data Preprocessing**: Collect and preprocess the data, including cleaning, tokenization, and feature extraction.
2. **Split Data**: Split the dataset into training and testing sets.
3. **Model Training**: Train the Naive Bayes classifier using the training data.
4. **Model Evaluation**: Evaluate the model's performance on the testing data using appropriate metrics such as accuracy, precision, recall, and F1-score.
5. **Model Tuning**: Experiment with different variants of Naive Bayes (e.g., Multinomial or Gaussian) and adjust hyperparameters as needed.
6. **Deployment**: Once satisfied with the model's performance, deploy it to make predictions on new, unseen data.

Conclusion

The Naive Bayes classifier is a simple yet effective machine learning algorithm with applications in text classification, spam detection, sentiment analysis, and more. While it has its limitations, its simplicity, efficiency, and suitability for text data make it a valuable tool in the machine learning toolkit. When properly implemented and tuned, Naive Bayes can provide meaningful insights and predictions for various real-world problems. However, it is essential to be mindful of the independence assumption and consider other algorithms when dealing with complex, interdependent features.

Implementation in python

Certainly, the Naïve Bayes Classifier is a popular choice for text classification and other machine learning tasks. Below is a simple implementation of the Naïve Bayes Classifier in Python using the scikit-learn library.

Step 1: Install scikit-learn

If you haven't already installed scikit-learn, you can do so using pip:

```
!pip install scikit-learn
```

Step 2: Import Necessary Libraries

```
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import accuracy_score, classification_report
from sklearn.model_selection import train_test_split
```

Step 3: Prepare Your Data

Assuming you have a dataset with text samples and corresponding labels (e.g., positive or negative sentiment), you should split the data into a training set and a test set. Here's an example:

```
# Sample data
texts = ["This is a positive review.", "Negative sentiment detected.", "A very positive experience.", "I didn't like this at all."]
# Corresponding labels (1 for positive, 0 for negative)
labels = [1, 0, 1, 0]
# Split the data into a training set and a test set
X_train, X_test, y_train, y_test = train_test_split(texts, labels, test_size=0.2, random_state=42)
```

Step 4: Feature Extraction

You need to convert the text data into numerical features. One common approach is to use the CountVectorizer, which counts the frequency of words in the text. Here's how to do it:

```
vectorizer = CountVectorizer()
X_train_vec = vectorizer.fit_transform(X_train)
X_test_vec = vectorizer.transform(X_test)
```

Step 5: Train the Naïve Bayes Classifier

Next, create and train the Naïve Bayes classifier. For text classification, the Multinomial Naïve Bayes classifier is commonly used:

```
clf = MultinomialNB()
clf.fit(X_train_vec, y_train)
```

Step 6: Make Predictions

Once the classifier is trained, you can use it to make predictions on new data:

```
y_pred = clf.predict(X_test_vec)
```

Step 7: Evaluate the Model
Evaluate the model's performance using appropriate metrics:
accuracy = accuracy_score(y_test, y_pred)
report = classification_report(y_test, y_pred) print(f"Accuracy: {accuracy}") print(report)
This code demonstrates a basic implementation of the Naïve Bayes Classifier in Python using scikit-learn. Depending on your specific task and dataset, you may need to fine-tune the pre-processing steps, hyper parameters, and model selection to achieve the best performance.

Questions
1. Implement in python program of the following problems using Bayes Theorem.
a) Of the students in the college, 60% of the students reside in the hostel and 40% of the students are day scholars. Previous year results report that 30% of all students who stay in the hostel scored A Grade and 20% of day scholars scored A grade. At the end of the year, one student is chosen at random and found that he/she has an A grade. What is the probability that the student is a hosteler?
b) Suppose you're testing for a rare disease, and you have the following information:

- The disease has a prevalence of 0.01 (1% of the population has the disease).
- The test is not perfect:
  - The test correctly identifies the disease (true positive) 99% of the time (sensitivity).
  - The test incorrectly indicates the disease (false positive) 2% of the time (1 - specificity).

Calculate the probability of having the disease given a positive test result using Bayes' theorem.

2. Develop a function python code for Naïve Bayes classifier from scratch without using scikit-learn library, to predict whether the buyer should buy computer or not. Consider a following sample training dataset stored in a CSV file containing information about following buyer conditions (such as “<=30,” “medium,” “Yes,” and “fair”) and whether the player played golf (“Yes” or “No”).

age	income	student	credit_rating	_com
<=30	high	no	fair	no
<=30	high	no	excellent	no
31...40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31...40	low	yes	excellent	yes
<=30	medium	no	fair	no
<=30	low	yes	fair	yes
>40	medium	yes	fair	yes
<=30	medium	yes	excellent	yes
31...40	medium	no	excellent	yes
31...40	high	yes	fair	yes
>40	medium	no	excellent	no

3. Write a Python function to implement the Naive Bayes classifier without using the scikit-learn library for the following sample training dataset stored as a .CSV file. Calculate the accuracy, precision, and recall for your train/test dataset.
a. Build a classifier that determines whether a text is about sports or not.
b. Determine which tag the sentence "A very close game" belongs to.

Text	Tag
"A great game"	Sports
"The election was over"	Not sports

"Very clean match"	Sports
"A clean but forgettable game"	Sports
"It was a close election"	Not sports

Additional Questions

1. Write a function python program to implement the naïve Bayesian classifier without using scikit-learn library for the following sample training data set stored as a .CSV file. Calculate the accuracy, precision, and recall for your train/test data set. To classify 'If the weather is sunny, then the Player should play or not'?

	Outlook	Play
0	Rainy	Yes
1	Sunny	Yes
2	Overcast	Yes
3	Overcast	Yes
4	Sunny	No
5	Rainy	Yes
6	Sunny	Yes
7	Overcast	Yes
8	Rainy	No
9	Sunny	No
10	Sunny	Yes
11	Rainy	No
12	Overcast	Yes
13	Overcast	Yes

2.Develop a function python code for Naïve Bayes classifier from scratch without using scikit-learn library, to predict whether the player should play golf based on weather conditions. Consider a following sample training dataset stored in a CSV file containing information about following weather conditions (such as "Sunny," "Hot," "High Humidity," and "Not Windy") and whether the player played golf ("Yes" or "No").

	WEATHER	TEMPERATURE	HUMIDITY	WINDY	PLAY GOLF
1	Rainy	Hot	High	False	No
2	Rainy	Hot	High	True	No
3	Overcast	Hot	High	False	Yes
4	Sunny	Mild	High	False	Yes
5	Sunny	Cool	Normal	False	Yes
6	Sunny	Cool	Normal	True	No

7	Overcast	Cool	Normal	True	Yes
8	Rainy	Mild	High	False	No
9	Rainy	Cool	Normal	False	Yes
10	Sunny	Mild	Normal	False	Yes
11	Rainy	Mild	Normal	True	Yes
12	Overcast	Mild	High	True	Yes
13	Overcast	Hot	Normal	False	Yes
14	Sunny	Mild	High	True	No

3. Write a function python program to implement the naïve Bayesian classifier without using scikit-learn library for a following sample training data set stored as a.CSV file. Calculate the accuracy, precision, and recall for your train/test data set.

	Text Documents	**Label**
1	I love this sandwich	pos
2	This is an amazing place	pos
3	I feel very good about these beers	pos
4	This is my best work	pos
5	What an awesome view	pos
6	I do not like this restaurant	neg
7	I am tired of this stuff	neg
8	I can't deal with this	neg
9	He is my sworn enemy	neg
10	My boss is horrible	neg
11	This is an awesome place	pos
12	I do not like the taste of this juice	neg
13	I love to dance	pos
14	I am sick and tired of this place	neg
15	What a great holiday	pos
16	That is a bad locality to stay	neg
17	We will have good fun tomorrow	pos
18	I went to my enemy's house today	neg

WEEK -08: K-Nearest Neighbour (K-NN) & ID-3 Decision Tree

Introduction

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique. K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm. K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems. K-NN is a **non-parametric algorithm**, which means it does not make any assumption on underlying data. It is also called a **lazy learner algorithm** because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset. KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

Example: Suppose, we have an image of a creature that looks similar to cat and dog, but we want to know either it is a cat or dog. So for this identification, we can use the KNN algorithm, as it works on a similarity measure. Our KNN model will find the similar features of the new data set to the cats and dogs images and based on the most similar features it will put it in either cat or dog category.

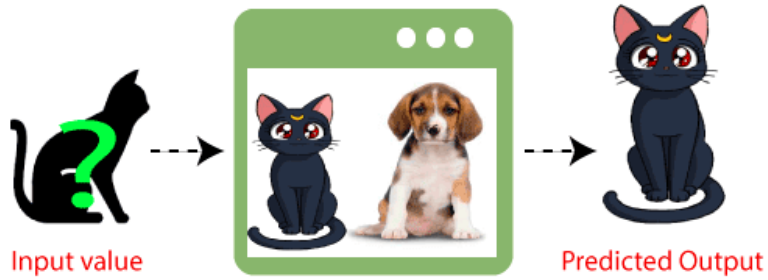
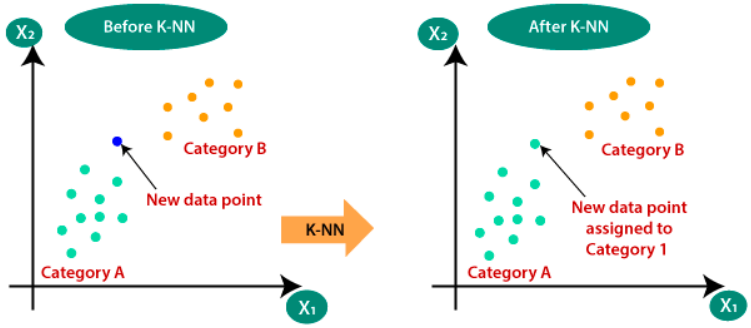


Figure 1 K-Nearest Neighbor Classifier

Need of a K-Nearest Neighbour Algorithm

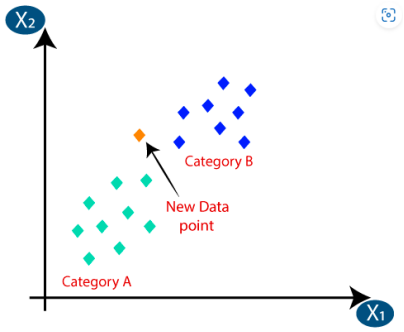
Suppose there are two categories, i.e., Category A and Category B, and we have a new data point x1, so this data point will lie in which of these categories. To solve this type of problem, we need a K-NN algorithm. With the help of K-NN, we can easily identify the category or class of a particular dataset. Consider the below diagram:



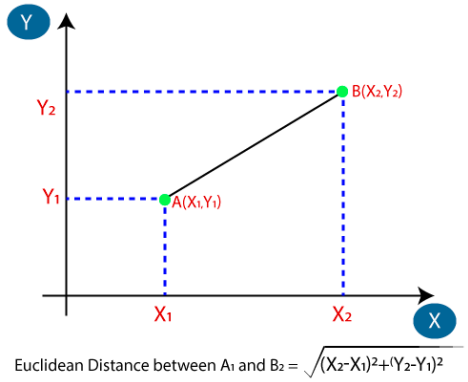
Working of a K-NN

Step-1: Select the number K of the neighbors
Step-2: Calculate the Euclidean distance of K number of neighbors
Step-3: Take the K nearest neighbors as per the calculated Euclidean distance.
Step-4: Among these k neighbors, count the number of the data points in each category.
Step-5: Assign the new data points to that category for which the number of the neighbor is maximum.
Step-6: Our model is ready.

Suppose we have a new data point and we need to put it in the required category. Consider the below image:



- Firstly, we will choose the number of neighbors, so we will choose the k=5.
- Next, we will calculate the Euclidean distance between the data points. The Euclidean distance is the distance between two points, which we have already studied in geometry. It can be calculated as:



By calculating the Euclidean distance we got the nearest neighbors, as three nearest neighbors in category A and two nearest neighbors in category B. Consider the below image:

As we can see the 3 nearest neighbors are from category A, hence this new data point must belong to category A.

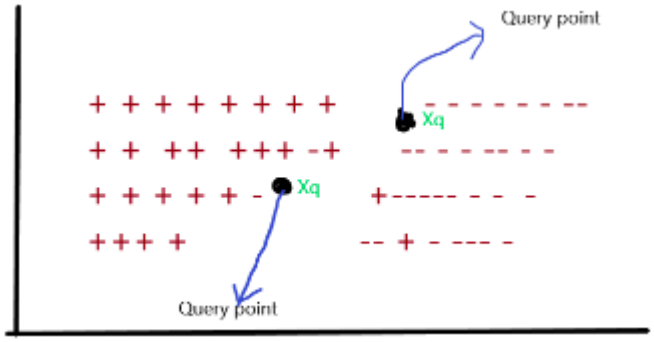
Selection of the value K in the K-NN Algorithm

Below are some points to remember while selecting the value of K in the K-NN algorithm:

- There is no particular way to determine the best value for "K", so we need to try some values to find the best out of them. The most preferred value for K is 5.
- A very low value for K such as K=1 or K=2, can be noisy and lead to the effects of outliers in the model.
- Large values for K are good, but it may find some difficulties.

Other Distance Measures used in K-NN

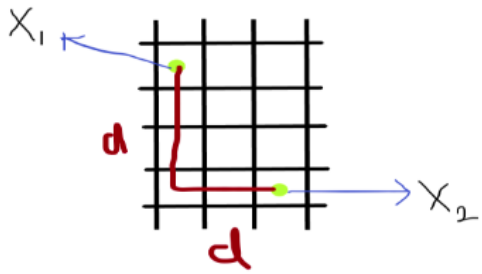
Nearest Neighbour:
let's take the simplest case of binary classification, suppose we have a group of +ve and -ve points in the dataset D such that the Xis belongs to the R-dim. data points and Yi are labels (+ve and -ve).



From the above image, you can conclude that there are several data points in 2 dim. Having the specific label, they are classified according to the +ve and -ve labels. If you noticed in the image there is one Query point referred to as **Xq** which has an unknown label. The surrounding points of **Xq** we considered as neighbours of Xq and the points which are close to the Xq are nearest neighbours. So how can we conclude that this point is nearest or not? It's by finding the distance b/w the points. So, here's the distance measures come existence.

We generally say that we will use distance to find the nearest neighbours of any query point Xq, but we still don't know how mathematically distance is measured between Xq and other nearest points? for further finding distance, we can't conclude that this is nearest or not.

In a theoretical manner, we can say that a distance measure is an objective score that summarizes the difference between two objects in a specific domain. There are several types of distance measures techniques but we only use some of them and they are listed below:
Manhatten Distance
This is the simplest way or technique to calculate the distance between two points, often called Taxicab distance or City Block distance, you can easily relate this with your daily life, If you start from somewhere and reached some destination so the Manhatten distance says that the distance between your starting point and the destination point. More mathematical we can say that It calculates the **absolute value** between two points. We calculate the distance exactly as the original path is we didn't take any diagonal or shortest path.
Let's take the geometric intuition for better understanding;

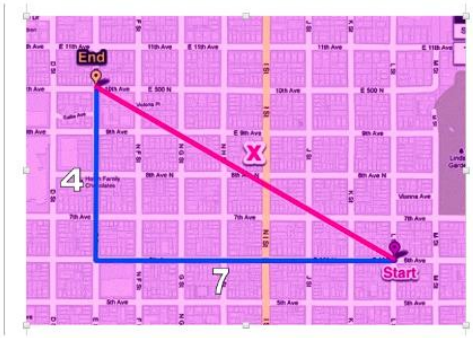


you can see in the image that the points **X1** and **X2** are 2-Dim vectors and having the coordinates same as before we discussed for euclidean distance, but here we can't calculate the distance as we calculate earlier apart from this we take the absolute value of the path from X1 to X2. In the image see that we take the path that actually covers from one point to another.

Manhatten Distance = sum for i to N sum || $X1_i$ – $X2_i$ ||
In mathematically we can write as :

$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

Note: Manhatten distance between two vectors or points is the L1 norm of two vector



As you see in the image the blue line represents the absolute path that the cab travel.
as Manhatten formula says: **distance = absolute sum ||xi-yi||**
distance = (7 + 4)
distance = 11
So, the absolute path the cab cover is 11.

Minkowski distance
Above we have discussed the L1 norm and L2 norm so this is the Lp norm of two vectors, More often we say that the Minkowski distance is the generalization or generalized form of the euclidean and Manhatten distance. Why do we call it to generalize? because we take both the distance technique and the new technique for finding the distance between vectors. It adds a parameter, called the "P", that allows different distance measures to be calculated.
Let's take it more mathematically, the equation of Minkowski distance is:

$$\|X_1 - X_2\| = \left(\sum_{i=1}^{d} |X_{1i} - X_{2i}|^p \right)^{1/p}$$

From the above equation you notice that the formula is the same as Euclidean distance but the change is that here we prefer the value of P, So if we take the P-value equals to 2 then it is euclidian distance and takes P-value equals to 1 then it is considered as Manhatten distance.
P = 1 => D = (∑i=1 to n(X1i−X2i)^1)$^{1/1}$
P = 2 => D = (∑i=1 to n(X1i−X2i)^2)$^{1/2}$
Note: **Minkowski distance between two vectors or points is the Lp norm of two vector.**

$$d(x, y) = \left(\sum_{i=1}^{n} |x_i - y_i|^c \right)^{\frac{1}{c}}$$

Example:
So, here we will take the same example as we take in the euclidean distance measures.
X1(x1, y1) = X1(3, 4)
X2(x2, y2) = X2(4, 7)
and we take the value of P = 4
distance = ((x2-x1)^4 + (y2-y1)^4)$^{1/4}$
distance = ((4-3)^4 + (7-4)^4)$^{1/4}$
distance = ((1)^4 + (3)^4)$^{1/4}$
distance = (1+81)$^{1/4}$
distance = (82)$^{1/4}$
distance = approx(2)

Pseudo Code of KNN

We can implement a KNN model by following the below steps:

1. Load the data
2. Initialise the value of k
3. For getting the predicted class, iterate from 1 to total number of training data points
    a. Calculate the distance between test data and each row of training dataset. Here we will use Euclidean distance as our distance metric since it's the most popular method. The other distance function or metrics that can be used are Manhattan distance, Minkowski distance, etc. If there are categorical variables, hamming distance can be used.
    b. Sort the calculated distances in ascending order based on distance values
    c. Get top k rows from the sorted array
    d. Get the most frequent class of these rows
    e. Return the predicted class

ID-3 Decision Tree

A decision tree is a structure that contains nodes (rectangular boxes) and edges(arrows) and is built from a dataset (table of columns representing features/attributes and rows corresponds to records). Each node is either used to make a decision (known as decision node) or represent an outcome (known as leaf node).

ID3 stands for Iterative Dichotomiser 3 and is named such because the algorithm iteratively (repeatedly) dichotomizes(divides) features into two or more groups at each step. Invented by Ross Quinlan, ID3 uses a top-down greedy approach to build a decision tree. In simple words, the top-down approach means that we start building the tree from the top and the greedy approach means that at each iteration we select the best feature at the present moment to create a node.

Metrics in ID3

ID3 uses **Information Gain** or just **Gain** to find the best feature. Information Gain calculates the reduction in the entropy and measures how well a given feature separates or classifies the target classes. The feature with the **highest Information Gain** is selected as the **best** one. **Entropy** is the measure of disorder and the Entropy of a dataset is the measure of disorder in the target feature of the dataset.

Information Gain= Entropy(S)- [(Weighted Avg) *Entropy(each feature)

Entropy(s)= -P(yes)log2 P(yes)- P(no) log2 P(no)

Where,

S= Total number of samples

P(yes)= probability of yes

P(no)= probability of no

ID3 Steps

1. Calculate the Information Gain of each feature.
2. Considering that all rows don't belong to the same class, split the dataset S into subsets using the feature for which the Information Gain is maximum.
3. Make a decision tree node using the feature with the maximum Information gain.
4. If all rows belong to the same class, make the current node as a leaf node with the class as its label.
5. Repeat for the remaining features until we run out of all features, or the decision tree has all leaf nodes.

Questions

Q-1. A Classification of Fruits

You are provided with a dataset of fruits. Each fruit is characterized by two features: weight (in grams) and sweetness level (on a scale of 1 to 10). You want to classify a new fruit as either an "Apple" or an "Orange" based on these features using the KNN algorithm.

Fruit ID	Weight (grams)	Sweetness Level	Label (Fruit Type)
1	180	7	Apple
2	200	6	Apple
3	150	4	Orange
4	170	5	Orange
5	160	6	Apple
6	140	3	Orange

Tasks:

1. Implement the KNN algorithm manually with k=3 to classify a new fruit with a weight of 165 grams and sweetness level of 5.5.
2. Calculate the Euclidean, Manhattan, and Minkowski distances between the new fruit and all the existing fruits in the dataset. Finally compare the calculated distances.
3. Based on the k-nearest neighbors, determine the label for the new fruit.
4. What is the effect of choosing different values of k (e.g., k=1, k=5) on the classification result?
5. Implement the above using function python program without using scikit learn library.
6. Plot the given samples, the Apple in Red color and the Orange in orange color. Also draw the decision boundary.

Q-1. B Classification of Fruits

Implement the Python code for Q-1. A using the ***scikit-learn*** library. Plot the given samples, using red for "Apple" and orange for "Orange." Also, plot the decision boundary. Calculate the distances using Euclidean, Manhattan, and Minkowski metrics, and compare the results.

Q-2. A Medical Diagnosis Decision

A dataset is provided to classify patients as "Healthy" or "Sick" based on their Age, Blood Pressure, and Cholesterol levels.

Patient ID	Age (years)	Blood Pressure (High/Low)	Cholesterol (High/Normal)	Diagnosis (Healthy/Sick)
1	30	High	High	Sick
2	45	Low	Normal	Healthy
3	50	High	High	Sick
4	35	Low	Normal	Healthy
5	60	High	High	Sick
6	55	Low	Normal	Healthy
7	40	High	High	Sick
8	25	Low	Normal	Healthy
9	65	High	High	Sick
10	45	Low	Normal	Healthy

Tasks:

1. Calculate the entropy for the target variable (Diagnosis).
2. Calculate the information gain for each feature (Age, Blood Pressure, Cholesterol).

3. Using the ID3 algorithm, decide which feature should be chosen as the root node for the decision tree.
4. Build the decision tree and explain the first few splits.
5. Predict whether a 50-year-old patient with low blood pressure and normal cholesterol is healthy or sick using the tree you built.
6. Implement the above using function python program without using scikit learn library.

Q-2. B Medical Diagnosis Decision

Implement the Python code for Q-2. A using the ***scikit-learn*** library. Using the ID3 algorithm, decide which feature should be chosen as the root node for the decision tree. Build the decision tree and explain the first few splits. Predict whether a 50-year-old patient with low blood pressure and normal cholesterol is healthy or sick using the tree you built.

Additional Questions

1. We have data from the questionnaires survey (to ask people opinion) and objective testing with two attributes (acid durability and strength) to classify whether a special paper tissue is good or not. Here is four training samples as follows. Apply the K-nearest neighbour's (KNN) algorithm when K=2, 3 and 4 to classify an instance (3, 7) as good or bad.

X1 = Acid Durability (seconds)	X2 = Strength (kg/square meter)	Y = Classification
7	7	Bad
7	4	Bad
3	4	Good
1	4	Good
4	5	Bad
3	5	Good
4	6	Bad
8	7	Bad
7	9	Good
8	8	Bad

Implement the above using python code without using scikit learn library. Plot the given samples Bad in Red color and Good in green color. Also draw the decision boundary. Calculate the desistance using Euclidean, Manhattan, and Minkowski and compare.

2. Implement the Question number 1 with using scikit learn library. Plot the given samples Bad in Red color and Good in green color. Also plot the decision boundary. Calculate the desistance using Euclidean, Manhattan, and Minkowski and compare the results.

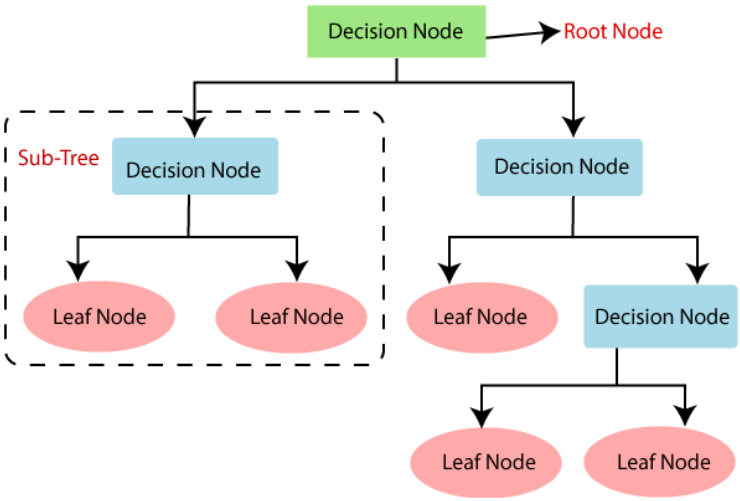
3.There is a Car manufacturer company that has manufactured a new SUV car. The company wants to give the ads to the users who are interested in buying that SUV. So for this problem, we have a dataset that contains multiple user's information through the social network. The dataset contains lots of information but the Estimated Salary and Age we will consider for the independent variable and the Purchased variable is for the dependent variable. Below is the dataset:

1. Apply the K-NN algorithm when K=2, 3 and 4 to classify purchased or not. Calculate the desistance using Euclidean, Manhattan, and Minkowski and compare.
2. Test your developed K-NN without and with using Scikit Learn Library.
3. Plot the Yellow points are for Purchased(1) and Green Points for not Purchased(0) variable.
4. Show the graph has to classify users in the correct categories, as most of the users who didn't buy the SUV are in the red region, and users who bought the SUV are in the green region.

User ID	Gender	Age	EstimatedSalary	Purchased
15624510	Male	19	19000	0
15810944	Male	35	20000	0
15668575	Female	26	43000	0
15603246	Female	27	57000	0
15804002	Male	19	76000	0
15728773	Male	27	58000	0
15598044	Female	27	84000	0
15694829	Female	32	150000	1
15600575	Male	25	33000	0
15727311	Female	35	65000	0
15570769	Female	26	80000	0
15606274	Female	26	52000	0
15746139	Male	20	86000	0
15704987	Male	32	18000	0
15628972	Male	18	82000	0
15697686	Male	29	80000	0
15733883	Male	47	25000	1
15617482	Male	45	26000	1
15704583	Male	46	28000	1
15621083	Female	48	29000	1
15649487	Male	45	22000	1
15736760	Female	47	49000	1

WEEK: 09 DECISION TREE [C4.5 AND CART]

- Decision Tree is a **Supervised learning technique** that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where **internal nodes represent the features of a dataset, branches represent the decision rules** and **each leaf node represents the outcome.**
- In a Decision tree, there are two nodes, which are the **Decision Node** and **Leaf Node.** Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.
- The decisions or the test are performed on the basis of features of the given dataset.
- ***It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.***
- It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.
- In order to build a tree, we use the **CART algorithm,** which stands for **Classification and Regression Tree algorithm.**
- A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees.
- Below diagram explains the general structure of a decision tree:



Why use Decision Trees?

There are various algorithms in Machine learning, so choosing the best algorithm for the given dataset and problem is the main point to remember while creating a machine learning model. Below are the two reasons for using the Decision tree:

- Decision Trees usually mimic human thinking ability while making a decision, so it is easy to understand.
- The logic behind the decision tree can be easily understood because it shows a tree-like structure.

Decision Tree Terminologies

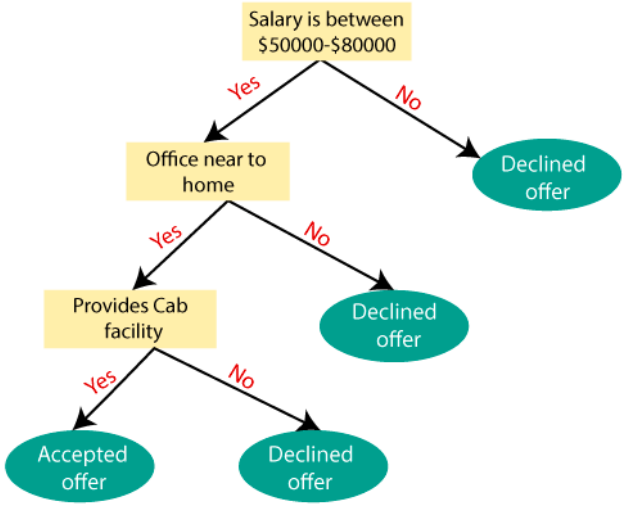
- **Root Node:** Root node is from where the decision tree starts. It represents the entire dataset, which further gets divided into two or more homogeneous sets.
- **Leaf Node:** Leaf nodes are the final output node, and the tree cannot be segregated further after getting a leaf node.
- **Splitting:** Splitting is the process of dividing the decision node/root node into sub-nodes according to the given conditions.
- **Branch/Sub Tree:** A tree formed by splitting the tree.
- **Pruning:** Pruning is the process of removing the unwanted branches from the tree.
- **Parent/Child node:** The root node of the tree is called the parent node, and other nodes are called the child nodes.

How does the Decision Tree algorithm Work?

In a decision tree, for predicting the class of the given dataset, the algorithm starts from the root node of the tree. This algorithm compares the values of root attribute with the record (real dataset) attribute and, based on the comparison, follows the branch and jumps to the next node.
For the next node, the algorithm again compares the attribute value with the other sub-nodes and move further. It continues the process until it reaches the leaf node of the tree. The complete process can be better understood using the below algorithm:

- **Step-1:** Begin the tree with the root node, says S, which contains the complete dataset.
- **Step-2:** Find the best attribute in the dataset using **Attribute Selection Measure (ASM).**
- **Step-3:** Divide the S into subsets that contains possible values for the best attributes.
- **Step-4:** Generate the decision tree node, which contains the best attribute.
- **Step-5:** Recursively make new decision trees using the subsets of the dataset created in step -3. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

Example: Suppose there is a candidate who has a job offer and wants to decide whether he should accept the offer or Not. So, to solve this problem, the decision tree starts with the root node (Salary attribute by ASM). The root node splits further into the next decision node (distance from the office) and one leaf node based on the corresponding labels. The next decision node further gets split into one decision node (Cab facility) and one leaf node. Finally, the decision node splits into two leaf nodes (Accepted offers and Declined offer). Consider the below diagram:



Attribute Selection Measures

While implementing a Decision tree, the main issue arises that how to select the best attribute for the root node and for sub-nodes. So, to solve such problems there is a technique which is called as **Attribute selection measure or ASM.** By this measurement, we can easily select the best attribute for the nodes of the tree. There are two popular techniques for ASM, which are:

- **Information Gain**
- **Gini Index**

***1. Information Gain*:**

- Information gain is the measurement of changes in entropy after the segmentation of a dataset based on an attribute.
- It calculates how much information a feature provides us about a class.
- According to the value of information gain, we split the node and build the decision tree.
- A decision tree algorithm always tries to maximize the value of information gain, and a node/attribute having the highest information gain is split first. It can be calculated using the below formula:

Information Gain= Entropy(S)- [(Weighted Avg) *Entropy(each feature)

Entropy: Entropy is a metric to measure the impurity in a given attribute. It specifies randomness in data. Entropy can be calculated as:

Entropy(s)= -P(yes)log2 P(yes)- P(no) log2 P(no)

Where,

- S= Total number of samples
- P(yes)= probability of yes

- P(no)= probability of no

2. Gini Index:

- Gini index is a measure of impurity or purity used while creating a decision tree in the CART(Classification and Regression Tree) algorithm.
- An attribute with the low Gini index should be preferred as compared to the high Gini index.
- It only creates binary splits, and the CART algorithm uses the Gini index to create binary splits.
- Gini index can be calculated using the below formula:

Gini Index= 1- $\sum_j P_j^2$

Pruning: Getting an Optimal Decision tree

Pruning is a process of deleting the unnecessary nodes from a tree in order to get the optimal decision tree.

A too-large tree increases the risk of overfitting, and a small tree may not capture all the important features of the dataset. Therefore, a technique that decreases the size of the learning tree without reducing accuracy is known as Pruning. There are mainly two types of tree **pruning** technology used:

- **Cost Complexity Pruning**
- **Reduced Error Pruning.**

C4.5, AND CART DECISION TREES

Entropy-It is used for checking the impurity or uncertainty present in the data. Entropy is used to evaluate the quality of a split. When entropy is zero the sample is completely homogeneous, meaning that each instance belongs to the same class and entropy is one when the sample is equally divided between different classes.

Formula of Entropy -

$$Entropy = \sum_{i=1}^{C} -p_i * \log_2(p_i)$$

Information Gain- Information gain indicates how much information a particular feature/ variable give us about the final outcome.

Formula of information gain-

$$Gain(T, X) = Entropy(T) - Entropy(T, X)$$

C4.5

C4.5 Decision Tree is a complicated Algorithm to understand. It does require a lot of background knowledge.

Information Entropy

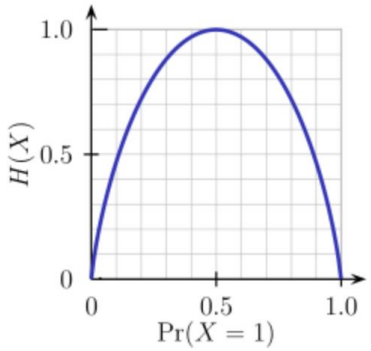
Information Entropy is the measure of impurity in a given example. Let's elaborate

Shanon entropy or self information:

$$-\Sigma_{i=1}^{n} P(x_i) \log_b(P(x_i))$$

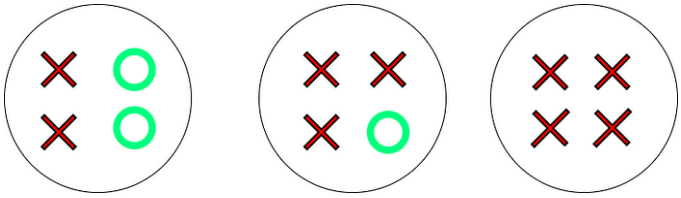
Where b is the base.

For a binary, the base would be 2 and the plot of the function would be as follows for percentages of 1 in the dataset:



So if the Probability of an event approaches 1 or 0, the Information Entropy tends to 0 as the output is more or less predictable.

In statistics, entropy measures the level of impurity(heterogeneity) in a dataset. A fully homogenous dataset has an entropy of 0 whereas a skewed dataset has an entropy closer to 1 as shown in the figure below:



Measure the impurity of a sample with Entropy

Information Gain

It is a parameter that is used to compute the change in entropy of a dataset before and after a transformation. Information Gain helps in feature selection. How it works is as follows.

A feature in a dataset is chosen and based on the value of the feature, the dataset is split into multiple smaller datasets. The change in the overall entropy from that of the parent to the average entropy of all the new child datasets is computed to be the Information Gain.

$$Information\ Gain = Entropy_{parent} - \sum Entropy_{child}$$

Formula for Information Gain

The higher the Information Gain the more accurately the feature divides the dataset as the resultant dataset are more homogenous with lower entropies. This indicates the feature has split(classified) the dataset accurately. On the other hand, if the Information Gain is low, it implies that the resultant datasets are more or less as heterogeneous as the parent dataset and so the feature does not provide much value.

Steps in algorithm:

· Check for the above base cases.

· For each attribute a, find the normalized information gain ratio from splitting on a.

· Let a_best be the attribute with the highest normalized information gain.

· Create a decision node that splits on a_best.

· Recurse on the sublists obtained by splitting on a_best, and add those nodes as children of node.

CART

CART stands for Classification And Regression Tree. It is a type of decision tree which can be used for both classification and regression tasks based on **non-parametric supervised learning** method. The following represents the algorithm steps. First and foremost, the data is split into training and test set.

- Take a feature K and split the training data set into two **subsets** based on some threshold of the feature, Tk. For example, if we are working through the IRIS data set, take a feature such as petal length, set the threshold as 2.25 cm. The question that would arise is how does the algorithm select K and Tk. The pair of K and Tk is selected in a way that purest **subsets** (weighted by their size) are created. This can also be called as the cost or loss function. The following cost function is optimized (minimized).

$$J\left(k, t_k\right) = \frac{m_{\text{left}}}{m} G_{\text{left}} + \frac{m_{\text{right}}}{m} G_{\text{right}}$$

$$\text{where } \begin{cases} G_{\text{left/right}} \text{ measures the impurity of the left/right subset,} \\ m_{\text{left/right}} \text{ is the number of instances in the left/right subset.} \end{cases}$$

- Once the training set is split into two subsets, the algorithm splits the subsets into another subsets using the same logic.
- The above split continues in a recursive manner until the maximum depth (hyperparameter max_depth) is reached or the algorithm is unable to find the split that further reduces the impurity.
- The following are few other hyperparameters which can be set before the algorithm is run for training. These are used to regularize the model.
  - min_samples_split: Minimum number of samples a node must have before it is split
  - min_samples_leaf: Minimum number of samples a leaf node must have
  - max_leaf_nodes: Maximum number of leaf nodes

CART decision tree is a **greedy algorithm** as it searches for optimum split right at the top most node without considering the possibility of lowest possible impurity several levels down. Note that greedy algorithms can result into a reasonably great solution but may not be optimal.

The following represents the CART cost function for regression. The cost function attempts to minimize the MSE (mean square error) for regression task. Recall that CART cost function for classification attempts to minimize impurity.

$$J\left(k, t_k\right)=\frac{m_{\text{left}}}{m} \mathrm{MSE}_{\text{left}}+\frac{m_{\text{right}}}{m} \mathrm{MSE}_{\text{right}} \quad \text{where} \begin{cases}\mathrm{MSE}_{\text{node}}=\sum_{i \in \text{node}}\left(\hat{y}_{\text{node}}-y^{(i)}\right)^2 \\ \hat{y}_{\text{node}}=\frac{1}{m_{\text{node}}} \sum_{i \in \text{node}} y^{(i)}\end{cases}$$

Decision Tree Pros

- Decision trees are easy to interpret and visualize.
- It can easily capture Non-linear patterns.
- It requires fewer data preprocessing from the user, for example, there is no need to normalize columns.
- It can be used for feature engineering such as predicting missing values, suitable for variable selection.
- The decision tree has no assumptions about distribution because of the non-parametric nature of the algorithm.

Decision Tree Cons

- Sensitive to noisy data. It can overfit noisy data.
- The small variation(or variance) in data can result in the different decision tree. This can be reduced by bagging and boosting algorithms.
- Decision trees are biased with imbalance dataset, so it is recommended that balance out the dataset before creating the decision tree.

Questions

1. Write a python function program to demonstrate the working of the decision tree based **C4.5** algorithms without using scikit-learn library. Use following data set for building the decision tree and apply this knowledge to classify a new sample.

The dataset has three attributes: Outlook (Sunny, Overcast, Rainy), Temperature, Humidity and Wind (Weak, Strong). The target attribute is Play Tennis (Yes/No).

Day	Outlook	Temp.	Humidity	Wind	Decision
1	Sunny	85	85	Weak	No
2	Sunny	80	90	Strong	No
3	Overcast	83	78	Weak	Yes
4	Rain	70	96	Weak	Yes
5	Rain	68	80	Weak	Yes
6	Rain	65	70	Strong	No
7	Overcast	64	65	Strong	Yes
8	Sunny	72	95	Weak	No
9	Sunny	69	70	Weak	Yes
10	Rain	75	80	Weak	Yes
11	Sunny	75	70	Strong	Yes
12	Overcast	72	90	Strong	Yes
13	Overcast	81	75	Weak	Yes
14	Rain	71	80	Strong	No

2. Write a python function program to demonstrate the working of the decision tree based **CART** algorithms without using scikit-learn library. Use Q. No. 1 data set for building the decision tree and apply this knowledge to classify a new sample.

The dataset has three attributes: Outlook (Sunny, Overcast, Rainy), Temperature, Humidity and Wind (Weak, Strong). The target attribute is Play Tennis (Yes/No).

3. Write a python function program to demonstrate the working of the decision tree based C4.5 and CART algorithms without and with using scikit-learn library. Using the following dataset, apply aforementioned algorithms. The attributes are **Income (Low, Medium, High)** and **Credit (Good, Bad)**, and the target is **Loan Approved (Yes/No)**.

Income	Credit	Loan Approved
Low	Good	Yes
Low	Bad	No
Medium	Good	Yes
Medium	Bad	Yes
High	Good	Yes
High	Bad	No

Additional Questions

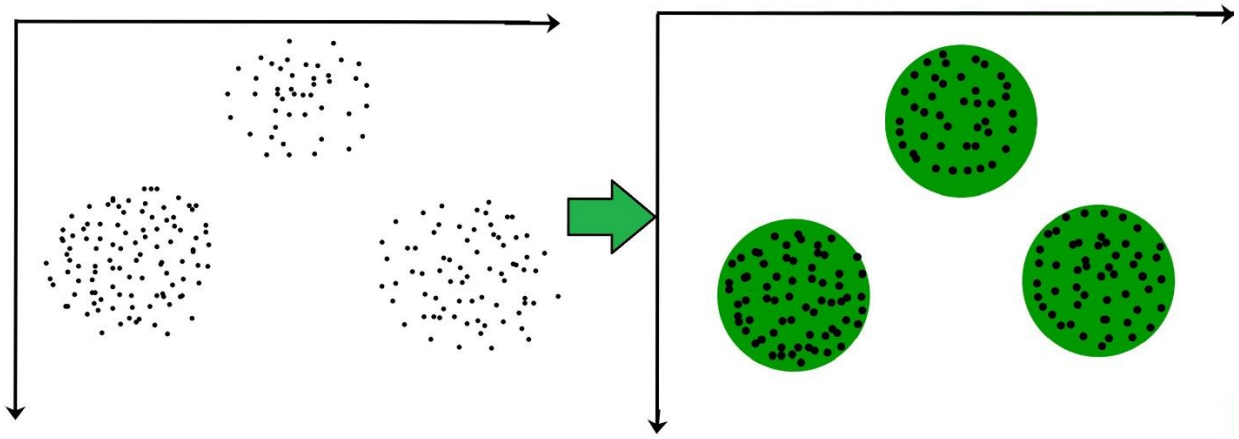
1. Write a program to demonstrate the working of the decision tree based **C4.5** algorithms with using scikit-learn library. Use an following data set (Q. No 1 under Questions) for building the decision tree and apply this knowledge to classify a new sample. Compare with and without using scikit-learn library, also the results. The dataset has three attributes: Outlook (Sunny, Overcast, Rainy), Temperature, Humidity and Wind (Weak, Strong). The target attribute is Play Tennis (Yes/No).
2. Write a program to demonstrate the working of the decision tree based **CART** algorithms with using scikit-learn library. Use an following data set (Q. No 1 under Questions) for building the decision tree and apply this knowledge to classify a new sample. Compare with and without using scikit-learn library, also the results. The dataset has three attributes: Outlook (Sunny, Overcast, Rainy), Temperature, Humidity and Wind (Weak, Strong). The target attribute is Play Tennis (Yes/No).

WEEK 10: DATA CLUSTERING K-means

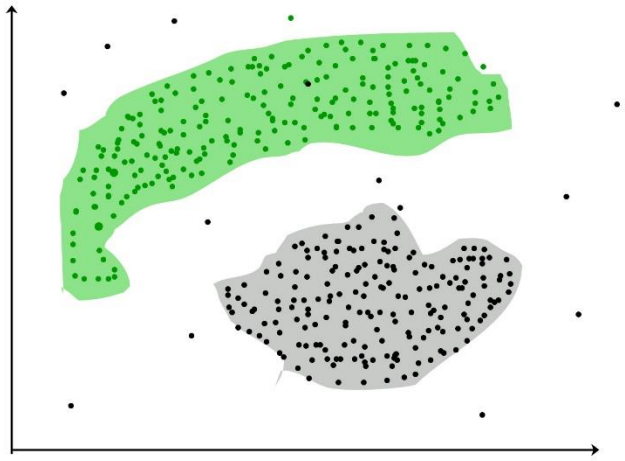
Introduction to Clustering: It is basically a type of *unsupervised learning method*. An unsupervised learning method is a method in which we draw references from datasets consisting of input data without labeled responses. Generally, it is used as a process to find meaningful structure, explanatory underlying processes, generative features, and groupings inherent in a set of examples.

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them.

For example The data points in the graph below clustered together can be classified into one single group. We can distinguish the clusters, and we can identify that there are 3 clusters in the below picture.



It is not necessary for clusters to be spherical as depicted below:



DBSCAN: Density-based Spatial Clustering of Applications with Noise

These data points are clustered by using the basic concept that the data point lies within the given constraint from the cluster center. Various distance methods and techniques are used for the calculation of the outliers.

Why Clustering?

Clustering is very much important as it determines the intrinsic grouping among the unlabelled data present. There are no criteria for good clustering. It depends on the user, and what criteria they may use which satisfy their need. For instance, we could be interested in finding representatives for homogeneous groups (data reduction), finding "natural clusters" and describing their unknown properties ("natural" data types), in finding useful and suitable groupings ("useful" data classes) or in finding unusual data objects (outlier detection). This algorithm must make some assumptions that constitute the similarity of points and each assumption make different and equally valid clusters.

Clustering Methods:

- **Density-Based Methods:** These methods consider the clusters as the dense region having some similarities and differences from the lower dense region of the space. These methods have good accuracy

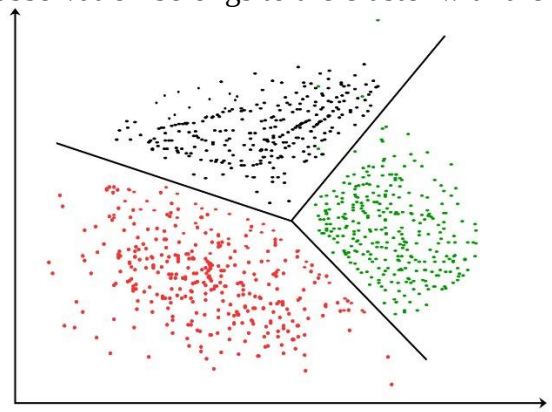
and the ability to merge two clusters. Example *DBSCAN (Density-Based Spatial Clustering of Applications with Noise), OPTICS (Ordering Points to Identify Clustering Structure)*, etc.

- **Hierarchical Based Methods:** The clusters formed in this method form a tree-type structure based on the hierarchy. New clusters are formed using the previously formed one. It is divided into two category
    - **Agglomerative** (bottom-up *approach*)
    - **Divisive** (top-down *approach*)

Examples *CURE (Clustering Using Representatives), BIRCH (Balanced Iterative Reducing Clustering and using Hierarchies)*, etc.

- **Partitioning Methods:** These methods partition the objects into k clusters and each partition forms one cluster. This method is used to optimize an objective criterion similarity function such as when the distance is a major parameter example *K-means, CLARANS (Clustering Large Applications based upon Randomized Search)*, etc.
- **Grid-based Methods:** In this method, the data space is formulated into a finite number of cells that form a grid-like structure. All the clustering operations done on these grids are fast and independent of the number of data objects example *STING (Statistical Information Grid), wave cluster, CLIQUE (CLustering In Quest)*, etc.

Clustering Algorithms: K-means clustering algorithm – It is the simplest unsupervised learning algorithm that solves clustering problem.K-means algorithm partitions n observations into k clusters where each observation belongs to the cluster with the nearest mean serving as a prototype of the cluster.



1. K-Means Clustering

K-Means Clustering is an unsupervised learning algorithm that is used to solve the clustering problems in machine learning or data science. In this topic, we will learn what is K-means clustering algorithm, how the algorithm works, along with the Python implementation of k-means clustering.

The goal of clustering is to divide the population or set of data points into a number of groups so that the data points within each group are more comparable to one another and different from the data points within the other groups. It is essentially a grouping of things based on how similar and different they are to one another.

We are given a data set of items, with certain features, and values for these features (like a vector). The task is to categorize those items into groups. To achieve this, we will use the K-means algorithm; an unsupervised learning algorithm. 'K' in the name of the algorithm represents the number of groups/clusters we want to classify our items into.

(It will help if you think of items as points in an n-dimensional space). The algorithm will categorize the items into k groups or clusters of similarity. To calculate that similarity, we will use the Euclidean distance as a measurement.

The algorithm works as follows:

1. First, we randomly initialize k points, called means or cluster centroids.
2. We categorize each item to its closest mean and we update the mean's coordinates, which are the averages of the items categorized in that cluster so far.

3. We repeat the process for a given number of iterations and at the end, we have our clusters.

The "points" mentioned above are called means because they are the mean values of the items categorized in them. To initialize these means, we have a lot of options. An intuitive method is to initialize the means at random items in the data set. Another method is to initialize the means at random values between the boundaries of the data set (if for a feature *x*, the items have values in [0,3], we will initialize the means with values for *x* at [0,3]).

The above algorithm in pseudocode is as follows:
Initialize k means with random values
--> For a given number of iterations:
--> Iterate through items:
--> Find the mean closest to the item by calculating the euclidean distance of the item with each of the means
--> Assign item to mean
--> Update mean by shifting it to the average of the items in that cluster

Example

Load the Datasets

```
from sklearn.datasets import load_digits
digits_data = load_digits().data
```

Each handwritten digit in the data is an array of color values of pixels of its image. For better understanding, let's print how the data of the first digit looks like and then display its's respective image.

```
import matplotlib.pyplot as plt
print(digits_data[0])
sample_digit = digits_data[0].reshape(8, 8)
plt.imshow(sample_digit)
plt.title("Digit image")
plt.show()
```

In the next step, we scale the data. Scaling is an optional yet very helpful technique for the faster processing of the model. In our model, we scale the pixel values which are typically between 0 – 255 to -1 – 1, easing the computation and avoiding super large numbers. Another point to consider is that a train test split is not required for this model as it is unsupervised learning with no labels to test. Then, we define the k value, which is 10 as we have 0-9 digits in our data. Also setting up the target variable.

```
from sklearn.preprocessing import scale
scaled_data = scale(digits_data)
print(scaled_data)
Y = load_digits().target
print(Y)
```

Defining k-means clustering:

Now we define the K-means cluster using the KMeans function from the sklearn module.

Method 1: Using a Random initial cluster.

- Setting the initial cluster points as random data points by using the '**init**' argument.
- The argument '**n_init**' is the number of iterations the k-means clustering should run with different initial clusters chosen at random, in the end, the clustering with the least total variance is considered'
- The random state is kept to 0 (any number can be given) to fix the same random initial clusters every time the code is run.

```
from sklearn.cluster import KMeans
k = 10
kmeans_cluster = KMeans(init = "random", n_clusters = k, n_init = 10, random_state = 0)
```

Method 2: Using k-means++

It is similar to method-1 however, it is not completely random, and chooses the initial clusters far away from each other. Therefore, it should require fewer iterations in finding the clusters when compared to the random initialization.

```
kmeans_cluster = KMeans(init="k-means++", n_clusters=k, n_init=10, random_state=0)
```

Model Evaluation

We will use scores like silhouette score, time taken to reach optimum position, v_measure and some other important metrics.

```
def bench_k_means(estimator, name, data):
        initial_time = time()
        estimator.fit(data)
        print("Initial-cluster: " + name)
        print("Time taken: {0:0.3f}".format(time() - initial_time))
        print("Homogeneity: {0:0.3f}".format( metrics.homogeneity_score(Y, estimator.labels_)))
        print("Completeness: {0:0.3f}".format( metrics.completeness_score(Y, estimator.labels_)))
        print("V_measure: {0:0.3f}".format( metrics.v_measure_score(Y, estimator.labels_)))
        print("Adjusted random: {0:0.3f}".format( metrics.adjusted_rand_score(Y, estimator.labels_)))
        print("Adjusted mutual info: {0:0.3f}".format(metrics.adjusted_mutual_info_score(Y, estimator.labels_)))
        print("Silhouette: {0:0.3f}".format(metrics.silhouette_score(data, estimator.labels_, metric='euclidean',
                                        sample_size=300)))
```

We will now use the above helper function to evaluate the performance of our k means algorithm.

```
kmeans_cluster = KMeans(init="random", n_clusters=k, n_init=10, random_state=0)
bench_k_means(estimator=kmeans_cluster, name="random", data=digits_data)
kmeans_cluster = KMeans(init="k-means++", n_clusters=k,n_init=10, random_state=0)
bench_k_means(estimator=kmeans_cluster, name="random", data=digits_data)
```

Visualizing the K-means clustering for handwritten data:

- Plotting the k-means cluster using the scatter function provided by the matplotlib module.
- Reducing the large dataset by using Principal Component Analysis (PCA) and fitting it to the previously defined k-means++ model.
- Plotting the clusters with different colors, a centroid was marked for each cluster.

```
from sklearn.decomposition import PCA
import numpy as np
# Reducing the dataset
pca = PCA(2)
reduced_data = pca.fit_transform(digits_data)
kmeans_cluster.fit(reduced_data)
# Calculating the centroids
centroids = kmeans_cluster.cluster_centers_
label = kmeans_cluster.fit_predict(reduced_data)
unique_labels = np.unique(label)
# plotting the clusters:
plt.figure(figsize=(8, 8))
for i in unique_labels:
        plt.scatter(reduced_data[label == i, 0], reduced_data[label == i, 1], label=i)
plt.scatter(centroids[:, 0], centroids[:, 1], marker='x', s=169, linewidths=3, color='k', zorder=10)
plt.legend()
plt.show()
```

Conclusion

From the graph, we can observe the clusters of the different digits are approximately separable from one another.

How To Make Clustering in Machine Learning

To cluster data in Scikit-Learn using Python, you must process the data, train multiple classification algorithms and evaluate each model to find the classification algorithm that is the best predictor for your data

1. **Load data**

You can load any labelled dataset that you want to predict on. For instance, you can use fetch_openml('mnist_784') on the Mnist dataset to practice.

2. **Explore the dataset**

Use python pandas functions such as df.describe() and df.isnull().sum() to find how your data need to be processed prior training

3. **Preprocess data**

Drop, fill or impute missing, or unwanted values from your dataset to make sure that you don't introduce errors or bias into your data. Use pandas get_dummies(), drop(), and fillna() functions alongside some sklearn's libraries such as SimpleImputer or OneHotEncoder to preprocess your data.

4. **Split data into training and testing dataset**

To be able to evaluate the accuracy of your models, split your data into training and testing sets using sklearn's train_test_split. This will allow to train your data on the training set and predict and evaluate on the testing set.

5. **Create a pipeline to train multiple clustering algorithms and hyper-parameters**

Run multiple algorithms, and for each algorithm, try various hyper-parameters. This will allow to find the best performing model and the best parameters for that model. Use GridSearchCV() and Pipeline to help you with these tasks

6. **Evaluate the machine learning model**

Evaluate the model on its precision with methods such as the homogeneity_score() and completeness_score() and evaluate elements such as the confusion_matrix() in Scikit-learn

Reference

https://medium.com/analytics-vidhya/how-to-determine-the-optimal-k-for-k-means-708505d204eb

https://www.analyticsvidhya.com/blog/2021/05/k-mean-getting-the-optimal-number-of-clusters/

https://www.geeksforgeeks.org/clustering-distance-measures/

Questions

1. Consider the following three variables for 20 different basketball players: points, assists, and rebounds. Perform k-means clustering manually with K=2, using Euclidean distance. Show the working for one iteration in your Lab Observation Book by using Euclidean distance.

points	assists	rebounds
18.0	3.0	15
19.0	4.0	14
14.0	5.0	10
14.0	4.0	8
11.0	7.0	14
20.0	8.0	13
28.0	7.0	9
30.0	6.0	5
31.0	9.0	4
35.0	12.0	11
33.0	14.0	6
25.0	9.0	5
25.0	4.0	3
27.0	3.0	8
29.0	4.0	12
30.0	12.0	7
19.0	15.0	6
23.0	11.0	5

Write a Python function (without using the scikit-learn library) to create a DataFrame containing the three variables (points, assists, and rebounds) for 20 different basketball players.
Apply the K-means algorithm to identify clusters with K=1, 2, K=3, and K=4, using distance formulas such as Euclidean distance, Manhattan distance, and Minkowski distance. Perform the following tasks:
a. Create a scatter plot of the data points in blue.
b. Plot the clusters with data points in different colors for K=1, 2, 3, and 4 in separate graphs.
c. Create a plot showing the number of clusters on the x-axis and the Sum of Squared Errors (SSE) on the y-axis. Compute SSE for all iterations. Show the table of given data points against SSE for every iteration and use the total sum of SSE in the graph of K vs. SSE .
d. Show the optimal value of K using the Elbow method and mark the same in the graph.

2. Redo 1(a)-1(d) using Manhattan distance.

3. Redo 1(a)-1(d) using Minkowski distance.

Additional questions

1.Write a Python function (with scikit-learn library) to create a DataFrame containing the three variables (points, assists, and rebounds) for 20 different basketball players.
Apply the K-means algorithm to identify clusters with K=1, 2, K=3, and K=4, using distance formulas such as Euclidean distance, Manhattan distance, and Minkowski distance. Perform the following tasks:
a. Create a scatter plot of the data points in blue.
b. Plot the clusters with data points in different colors for K=1, 2, 3, and 4 in separate graphs.
c. Create a plot showing the number of clusters on the x-axis and the Sum of Squared Errors (SSE) on the y-axis. Compute SSE for all iterations. Show the table of given data points against SSE for every iteration and use the total sum of SSE in the graph of K vs. SSE .
d. Show the optimal value of K using the Elbow method and mark the same in the graph.

2. Redo 1(a)-1(d) using Manhattan distance.

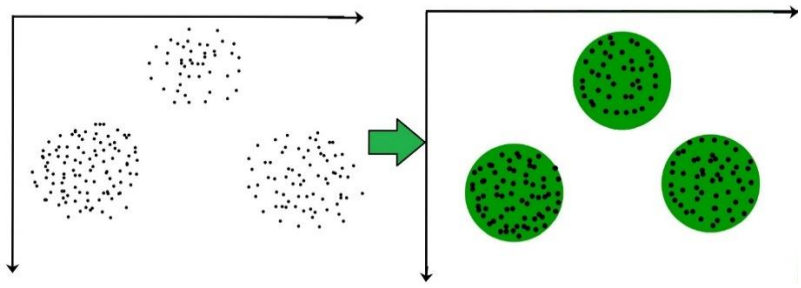
3. Redo 1(a)-1(d) using Minkowski distance.

WEEK 10: DATA CLUSTERING-K-means

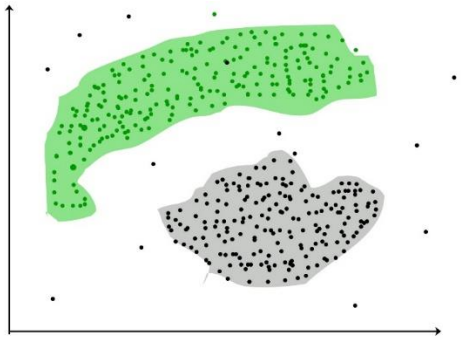
Introduction to Clustering: It is basically a type of *unsupervised learning method*. An unsupervised learning method is a method in which we draw references from datasets consisting of input data without labeled responses. Generally, it is used as a process to find meaningful structure, explanatory underlying processes, generative features, and groupings inherent in a set of examples.

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them.

For example The data points in the graph below clustered together can be classified into one single group. We can distinguish the clusters, and we can identify that there are 3 clusters in the below picture.



It is not necessary for clusters to be spherical as depicted below:



DBSCAN: Density-based Spatial Clustering of Applications with Noise

These data points are clustered by using the basic concept that the data point lies within the given constraint from the cluster center. Various distance methods and techniques are used for the calculation of the outliers.

Why Clustering?

Clustering is very much important as it determines the intrinsic grouping among the unlabelled data present. There are no criteria for good clustering. It depends on the user, and what criteria they may use which satisfy their need. For instance, we could be interested in finding representatives for homogeneous groups (data reduction), finding "natural clusters" and describing their unknown properties ("natural" data types), in finding useful and suitable groupings ("useful" data classes) or in finding unusual data objects (outlier detection). This algorithm must make some assumptions that constitute the similarity of points and each assumption make different and equally valid clusters.

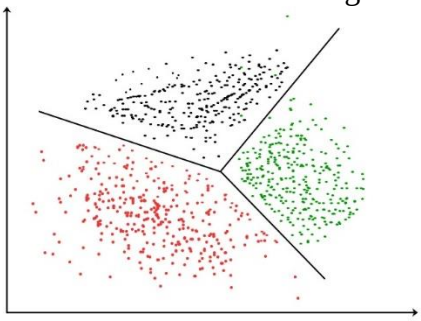
Clustering Methods:

- **Density-Based Methods:** These methods consider the clusters as the dense region having some similarities and differences from the lower dense region of the space. These methods have good accuracy and the ability to merge two clusters. Example *DBSCAN (Density-Based Spatial Clustering of Applications with Noise), OPTICS (Ordering Points to Identify Clustering Structure),* etc.
- **Hierarchical Based Methods:** The clusters formed in this method form a tree-type structure based on the hierarchy. New clusters are formed using the previously formed one. It is divided into two category
    - **Agglomerative** (bottom-up *approach*)
    - **Divisive** (top-down *approach*)

Examples *CURE (Clustering Using Representatives), BIRCH (Balanced Iterative Reducing Clustering and using Hierarchies),* etc.

- **Partitioning Methods:** These methods partition the objects into k clusters and each partition forms one cluster. This method is used to optimize an objective criterion similarity function such as when the distance is a major parameter example *K-means, CLARANS (Clustering Large Applications based upon Randomized Search)*, etc.
- **Grid-based Methods:** In this method, the data space is formulated into a finite number of cells that form a grid-like structure. All the clustering operations done on these grids are fast and independent of the number of data objects example *STING (Statistical Information Grid), wave cluster, CLIQUE (CLustering In Quest)*, etc.

Clustering Algorithms: K-means clustering algorithm – It is the simplest unsupervised learning algorithm that solves clustering problem.K-means algorithm partitions n observations into k clusters where each observation belongs to the cluster with the nearest mean serving as a prototype of the cluster.

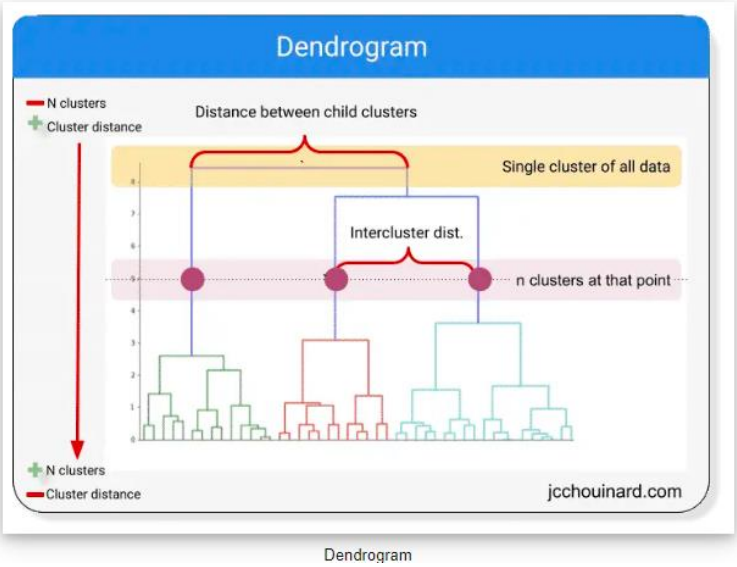


Hierarchical Clustering

Hierarchical clustering algorithm works by starting with 1 cluster per data point and merging the clusters together untilthe optimal clustering is met.

1. Having 1 cluster for each data point
2. Defining new cluster centers using the mean of X and Y coordinates
3. Combining clusters centers closest to each other
4. Finding new cluster centers based on the mean
5. Repeating until optimal number of clusters is met

The image below represents a **dendrogram** that can be used to visualize hierarchical clustering. Starting with 1 cluster per data point at the bottom and merging the closest clusters at each iteration, ending up with a single cluster for the entire dataset.



Dendrogram

Some examples of hierarchical clustering algorithms are:

- hirearchy from SciPy's scipy.cluster

Hierarchical Clustering in Python Example

```
import matplotlib.pyplot as plt
import numpy as np
from numpy.random import rand
import pandas as pd
import seaborn as sns
```

```
from scipy.cluster.vq import whiten
from scipy.cluster.hierarchy import fcluster, linkage

# Generate initial data
data = np.vstack(((rand(30,2)+1), (rand(30,2)+2.5),  (rand(30,2)+4)  ))

# standardize (normalize) the features
data = whiten(data)

# Compute the distance matrix
matrix = linkage(  data,  method='ward',  metric='euclidean'  )

# Assign cluster labels
labels = fcluster(  matrix, 3, criterion='maxclust'  )
# Create DataFrame
df = pd.DataFrame(data, columns=['x','y'])
df['labels'] = labels

# Plot Clusters
sns.scatterplot(  x='x',  y='y', hue='labels',  data=df )
plt.title('Hierachical Clustering with SciPy')
plt.show()
```

How To Make Clustering in Machine Learning

To cluster data in Scikit-Learn using Python, you must process the data, train multiple classification algorithms and evaluate each model to find the classification algorithm that is the best predictor for your data

1. **Load data**

You can load any labelled dataset that you want to predict on. For instance, you can use fetch_openml('mnist_784') on the Mnist dataset to practice.

2. **Explore the dataset**

Use python pandas functions such as df.describe() and df.isnull().sum() to find how your data need to be processed prior training

3. **Preprocess data**

Drop, fill or impute missing, or unwanted values from your dataset to make sure that you don't introduce errors or bias into your data. Use pandas get_dummies(), drop(), and fillna() functions alongside some sklearn's libraries such as SimpleImputer or OneHotEncoder to preprocess your data.

4. **Split data into training and testing dataset**

To be able to evaluate the accuracy of your models, split your data into training and testing sets using sklearn's train_test_split. This will allow to train your data on the training set and predict and evaluate on the testing set.

5. **Create a pipeline to train multiple clustering algorithms and hyper-parameters**

Run multiple algorithms, and for each algorithm, try various hyper-parameters. This will allow to find the best performing model and the best parameters for that model. Use GridSearchCV() and Pipeline to help you with these tasks

6. **Evaluate the machine learning model**

Evaluate the model on its precision with methods such as the homogeneity_score() and completeness_score() and evaluate elements such as the confusion_matrix() in Scikit-learn

Reference

https://www.geeksforgeeks.org/ml-types-of-linkages-in-clustering/

refer: Choosing the right linkage method for hierarchical clustering

https://stats.stackexchange.com/questions/195446/choosing-the-right-linkage-method-for-hierarchical-clustering

https://statsandr.com/blog/clustering-analysis-k-means-and-hierarchical-clustering-by-hand-and-in-r/#data-1

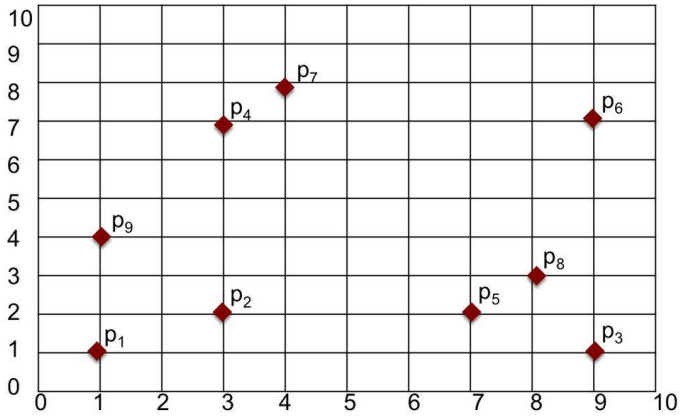
https://online.stat.psu.edu/stat505/lesson/14/14.4

Questions

1. For the following 1-dimensional data points apply agglomerative hierarchical clustering to build the dendrogram. Construct the proximity matrix (distance matrix). Merge the clusters using the single linkage (min distance) and update the proximity matrix accordingly. Clearly show the proximity matrix corresponding to each iteration of the algorithm.

18, 22, 25, 27, 42, 43

2. Consider the following data set and apply the hierarchical data-clustering algorithm, to identify the clusters. Solve it manually by considering all linkage functions (Single, Complete and Average) using Euclidean distance.



Point	x	y
p_1	1	1
p_2	3	2
p_3	9	1
p_4	3	7
p_5	7	2
p_6	9	7
p_7	4	8
p_8	8	3
p_9	1	4

3. Consider the above-mentioned data set in Q no 1 and apply the hierarchical data-clustering algorithm, to identify the clusters. Write a Python function (without using the **scikit-learn library**) to do the following:
a. Plot a scatter graph of given data points.
b. Display the proximity matrix using Euclidean distance, Manhattan distance, and Minkowski distance.
c. Plot the dendrogram for single, complete and average linkage methods.

Additional Questions
Consider the above-mentioned data set in Q no 1 and apply the hierarchical data-clustering algorithm, to identify the clusters. Write a Python function (with **scikit-learn library**) to do the following:
a. Plot a scatter graph of given data points.
b. Display the proximity matrix using Euclidean distance, Manhattan distance, and Minkowski distance.
c. Plot the dendrogram for single, complete and average linkage methods.

WEEK-12 SUPPORT VECTOR MACHINE

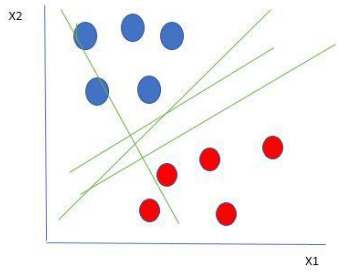
Support Vector Machine (SVM) is a powerful machine learning algorithm used for linear or nonlinear classification, regression, and even outlier detection tasks. SVMs can be used for a variety of tasks, such as text classification, image classification, spam detection, handwriting identification, gene expression analysis, face detection, and anomaly detection. SVMs are adaptable and efficient in a variety of applications because they can manage high-dimensional data and nonlinear relationships.

SVM algorithms are very effective as we try to find the maximum separating hyperplane between the different classes available in the target feature.

Support Vector Machine

Support Vector Machine (SVM) is a supervised machine learning algorithm used for both classification and regression. Though we say regression problems as well it's best suited for classification. The main objective of the SVM algorithm is to find the optimal hyperplane in an N-dimensional space that can separate the data points in different classes in the feature space. The hyperplane tries that the margin between the closest points of different classes should be as maximum as possible. The dimension of the hyperplane depends upon the number of features. If the number of input features is two, then the hyperplane is just a line. If the number of input features is three, then the hyperplane becomes a 2-D plane. It becomes difficult to imagine when the number of features exceeds three.

Let's consider two independent variables x_1, x_2, and one dependent variable which is either a blue circle or a red circle.

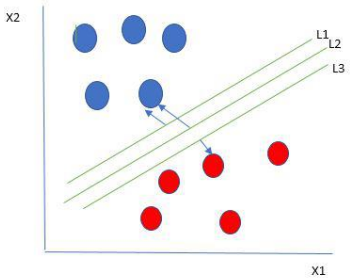


Linearly Separable Data points

From the figure above it's very clear that there are multiple lines (our hyperplane here is a line because we are considering only two input features x_1, x_2) that segregate our data points or do a classification between red and blue circles. So how do we choose the best line or in general the best hyperplane that segregates our data points?

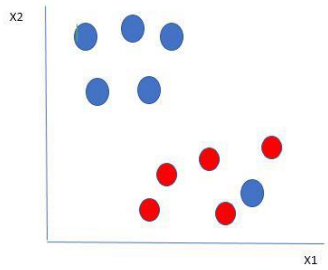
How does SVM work?

One reasonable choice as the best hyperplane is the one that represents the largest separation or margin between the two classes.



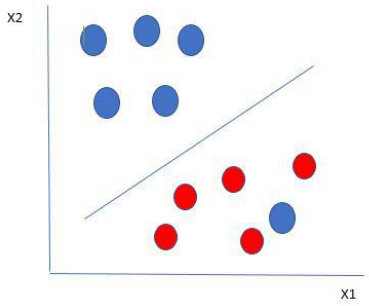
Multiple hyperplanes separate the data from two classes

So we choose the hyperplane whose distance from it to the nearest data point on each side is maximized. If such a hyperplane exists it is known as the maximum-margin hyperplane/hard margin. So from the above figure, we choose L2. Let's consider a scenario like shown below



Selecting hyperplane for data with outlier

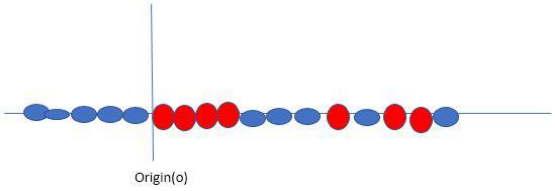
Here we have one blue ball in the boundary of the red ball. So how does SVM classify the data? It's simple! The blue ball in the boundary of red ones is an outlier of blue balls. The SVM algorithm has the characteristics to ignore the outlier and finds the best hyperplane that maximizes the margin. SVM is robust to outliers.



Hyperplane which is the most optimized one

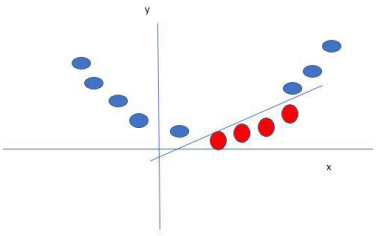
So in this type of data point what SVM does is, finds the maximum margin as done with previous data sets along with that it adds a penalty each time a point crosses the margin. So the margins in these types of cases are called soft margins. When there is a soft margin to the data set, the SVM tries to minimize *(1/margin+Λ(∑penalty))*. Hinge loss is a commonly used penalty. If no violations no hinge loss.If violations hinge loss proportional to the distance of violation.

Till now, we were talking about linearly separable data(the group of blue balls and red balls are separable by a straight line/linear line). What to do if data are not linearly separable?



Original 1D dataset for classification

Say, our data is shown in the figure above. SVM solves this by creating a new variable using a kernel. We call a point x_i on the line and we create a new variable y_i as a function of distance from origin o.so if we plot this we get something like as shown below



Mapping 1D data to 2D to become able to separate the two classes

In this case, the new variable y is created as a function of distance from the origin. A non-linear function that creates a new variable is referred to as a kernel.

Support Vector Machine Terminology

Hyperplane: Hyperplane is the decision boundary that is used to separate the data points of different classes in a feature space. In the case of linear classifications, it will be a linear equation i.e. wx+b = 0.

Support Vectors: Support vectors are the closest data points to the hyperplane, which makes a critical role in deciding the hyperplane and margin.

Margin: Margin is the distance between the support vector and hyperplane. The main objective of the support vector machine algorithm is to maximize the margin. The wider margin indicates better classification performance.
Kernel: Kernel is the mathematical function, which is used in SVM to map the original input data points into high-dimensional feature spaces, so, that the hyperplane can be easily found out even if the data points are not linearly separable in the original input space. Some of the common kernel functions are linear, polynomial, radial basis function(RBF), and sigmoid.
Hard Margin: The maximum-margin hyperplane or the hard margin hyperplane is a hyperplane that properly separates the data points of different categories without any misclassifications.
Soft Margin: When the data is not perfectly separable or contains outliers, SVM permits a soft margin technique. Each data point has a slack variable introduced by the soft-margin SVM formulation, which softens the strict margin requirement and permits certain misclassifications or violations. It discovers a compromise between increasing the margin and reducing violations.
C: Margin maximisation and misclassification fines are balanced by the regularisation parameter C in SVM. The penalty for going over the margin or misclassifying data items is decided by it. A stricter penalty is imposed with a greater value of C, which results in a smaller margin and perhaps fewer misclassifications.
Hinge Loss: A typical loss function in SVMs is hinge loss. It punishes incorrect classifications or margin violations. The objective function in SVM is frequently formed by combining it with the regularisation term.
Dual Problem: A dual Problem of the optimisation problem that requires locating the Lagrange multipliers related to the support vectors can be used to solve SVM. The dual formulation enables the use of kernel tricks and more effective computing.

Mathematical intuition of Support Vector Machine
Consider a binary classification problem with two classes, labeled as +1 and -1. We have a training dataset consisting of input feature vectors X and their corresponding class labels Y.
The equation for the linear hyperplane can be written as:

$$w^T x + b = 0$$

The vector W represents the normal vector to the hyperplane. i.e the direction perpendicular to the hyperplane. The parameter b in the equation represents the offset or distance of the hyperplane from the origin along the normal vector w.
The distance between a data point x_i and the decision boundary can be calculated as:

$$d_i = \frac{w^T x_i + b}{||w||}$$

where ||w|| represents the Euclidean norm of the weight vector w. Euclidean norm of the normal vector W
For Linear SVM classifier :

$$\hat{y} = \begin{cases} 1 & : \ w^T x + b \geq 0 \\ 0 & : \ w^T x + b < 0 \end{cases}$$

Optimization:
For Hard margin linear SVM classifier:

$$\underset{w,b}{\text{minimize}} \frac{1}{2} w^T w = \underset{W,b}{\text{minimize}} \frac{1}{2} \|w\|^2$$
$$\text{subject to } y_i(w^T x_i + b) \geq 1 \ for \ i = 1, 2, 3, \cdots, m$$

The target variable or label for the i^{th} training instance is denoted by the symbol t_i in this statement. And t_i=-1 for negative occurrences (when y_i= 0) and t_i=1positive instances (when y_i = 1) respectively. Because we require the decision boundary that satisfy the constraint: $t_i(w^T x_i + b) \geq 1$
For Soft margin linear SVM classifier:

$$\underset{w,b}{\text{minimize}} \ \frac{1}{2} w^T w + C \sum_{i=1}^{m} \zeta_i$$
$$\text{subject to } y_i(w^T x_i + b) \geq 1 - \zeta_i \ \ and \ \zeta_i \geq 0 \ \ for \ i = 1, 2, 3, \cdots, m$$

Dual Problem: A dual Problem of the optimisation problem that requires locating the Lagrange multipliers related to the support vectors can be used to solve SVM. The optimal Lagrange multipliers α(i) that maximize the following dual objective function

$$\underset{\alpha}{\text{maximize}} : \frac{1}{2} \sum_{i \to m} \sum_{j \to m} \alpha_i \alpha_j t_i t_j K(x_i, x_j) - \sum_{i \to m} \alpha_i$$

where,
α_i is the Lagrange multiplier associated with the ith training sample.
$K(x_i, x_j)$ is the kernel function that computes the similarity between two samples x_i and x_j. It allows SVM to handle nonlinear classification problems by implicitly mapping the samples into a higher-dimensional feature space.
The term $\sum \alpha_i$ represents the sum of all Lagrange multipliers.
The SVM decision boundary can be described in terms of these optimal Lagrange multipliers and the support vectors once the dual issue has been solved and the optimal Lagrange multipliers have been discovered. The training samples that have i > 0 are the support vectors, while the decision boundary is supplied by:

$$w = \sum_{i \to m} \alpha_i t_i K(x_i, x) + b$$
$$t_i(w^T x_i - b) = 1 \iff b = w^T x_i - t_i$$

Types of Support Vector Machine
Based on the nature of the decision boundary, Support Vector Machines (SVM) can be divided into two main parts:
Linear SVM: Linear SVMs use a linear decision boundary to separate the data points of different classes. When the data can be precisely linearly separated, linear SVMs are very suitable. This means that a single straight line (in 2D) or a hyperplane (in higher dimensions) can entirely divide the data points into their respective classes. A hyperplane that maximizes the margin between the classes is the decision boundary.
Non-Linear SVM: Non-Linear SVM can be used to classify data when it cannot be separated into two classes by a straight line (in the case of 2D). By using kernel functions, nonlinear SVMs can handle nonlinearly separable data. The original input data is transformed by these kernel functions into a higher-dimensional feature space, where the data points can be linearly separated. A linear SVM is used to locate a nonlinear decision boundary in this modified space.
Popular kernel functions in SVM
The SVM kernel is a function that takes low-dimensional input space and transforms it into higher-dimensional space, ie it converts nonseparable problems to separable problems. It is mostly useful in non-linear separation problems. Simply put the kernel, does some extremely complex data transformations and then finds out the process to separate the data based on the labels or outputs defined.

$$\text{Linear} : K(w, b) = w^T x + b$$
$$\text{Polynomial} : K(w, x) = (\gamma w^T x + b)^N$$
$$\text{Gaussian RBF}: K(w, x) = \exp(-\gamma ||x_i - x_j||^n$$
$$\text{Sigmoid} : K(x_i, x_j) = \tanh(\alpha x_i^T x_j + b)$$

Advantages of SVM
- Effective in high-dimensional cases.
- Its memory is efficient as it uses a subset of training points in the decision function called support vectors.
- Different kernel functions can be specified for the decision functions and its possible to specify custom kernels.

SVM implementation in Python

Predict if cancer is Benign or malignant. Using historical data about patients diagnosed with cancer enables doctors to differentiate malignant cases and benign ones are given independent attributes.

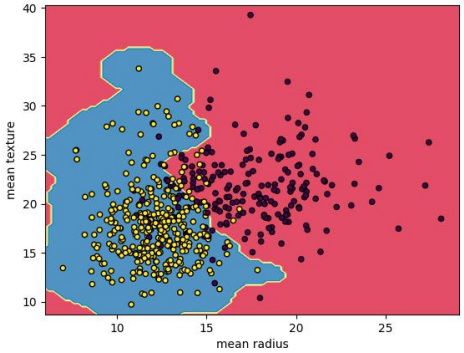
Steps

- Load the breast cancer dataset from sklearn.datasets
- Separate input features and target variables.
- Build and train the SVM classifiers using RBF kernel.
- Plot the scatter plot of the input features.
- Plot the decision boundary.

Python Code:

```
# Load the important packages
from sklearn.datasets import load_breast_cancer
import matplotlib.pyplot as plt
from sklearn.inspection import DecisionBoundaryDisplay
from sklearn.svm import SVC

# Load the datasets
cancer = load_breast_cancer()
X = cancer.data[:, :2]
y = cancer.target

#Build the model
svm = SVC(kernel="rbf", gamma=0.5, C=1.0)
# Trained the model
svm.fit(X, y)

# Plot Decision Boundary
DecisionBoundaryDisplay.from_estimator(
        svm,
        X,
        response_method="predict",
        cmap=plt.cm.Spectral,
        alpha=0.8,
        xlabel=cancer.feature_names[0],
        ylabel=cancer.feature_names[1],
    )

# Scatter plot
plt.scatter(X[:, 0], X[:, 1],
            c=y,
            s=20, edgecolors="k")
plt.show()
```

Output:



Questions

1.Use the given code and modify for IRIS dataset. Implement the SVM classifier in Python (make use of scikit-learn library). Apply the linear kernel function. Plot the scatter plot of the input features. Plot the decision boundary.

2. Construct a simple SVM classifier that separates the two classes:

Positively labeled data points: (4, 1), (4, -1), (6, 0)

Negatively labeled data points: (1, 0), (0, 1), (0, -1)

For all negatively labeled points, the output is -1, and for all positively labeled points, the output is 1. Implement the python function program to draw the hyperplane that separates the two classes using scikit-learn library. Plot the scatter plot of the input features.

3. Solve Qn (2) manually in your Observation book and implement without using scikit-learn library.

4. Develop a Python function program to sketch the hyperplane $1 + 2X1 + 3X2 = 0$ without using scikit-learn library. Indicate the set of points for which $1 + 3X1 - X2 > 0$, $1 + 3X1 - X2 < 0$ and $1 + 3X1 - X2=0$. Take x1, x2 ∈ (-10, 10). Plot the graph for every +/- 1 increment. Find the slope and intercept. Solve manually in your Observation book. Compare your results with manual results.

5. Given two hyperplanes for SVM classifier 1 and SVM classifier 2, find the best hyperplane corresponding to the classifier:

a. $5+2x_1+5x_2$
b. $5+20x_1+50x_2$

Implement Python function program to draw the hyperplane that separates the two classes (without scikit-learn library). Plot the scatter plot of the input features. Indicate the set of points for which
5+2*x1+ 5*x2 > 0 and
5+2*x1+ 5*x2 <0,
5+2*x1+ 5*x2 = 0
On the same plot, Indicate the set of points for which
5+20*x1+50*x2 > 0 and
5+20*x1+50*x2 < 0,
5+20*x1+50*x2 =0.
Find the slope and intercept. Solve manually in your Observation book. Compare your results with manual results.

Additional Questions

1. Given two hyperplanes for SVM classifier 1 and SVM classifier 2, find the best hyperplane corresponding to the classifier:

a. $5+2x_1+5x_2$
b. $5+20x_1+50x_2$

Implement the python function program to draw the hyperplane that separates the two classes (make use of scikit-learn libray). Plot the scatter plot of the input features. Indicate the set of points for which
5+2*x1+ 5*x2 > 0 and
5+2*x1+ 5*x2 <0,
5+2*x1+ 5*x2 = 0
On the same plot, Indicate the set of points for which
5+20*x1+50*x2 > 0 and
5+20*x1+50*x2 < 0,
5+20*x1+50*x2 =0.

2. Use the IRIS dataset, implement the SVM classifier in Python (make use of scikit-learn library), to do the following.

a. Apply the kernel functions such as linear, polynomial, Radial basis functions and Sigmoid.
b. Plot the scatter plot of the input features.
c. Plot the decision boundary.